

# Rebuttal Report

## Review of Principal Components Analysis of Data and Review of Inferences about Presence of Biomarkers in the Population of Animals from the Illinois River Watershed

**Prepared for:**

Tyson Foods, Inc.  
Tyson Poultry, Inc.  
Tyson Chicken, Inc.  
Cobb-Vantress, Inc.  
Cal-Maine Foods, Inc.  
Cal-Maine Farms, Inc.  
Cargill, Inc.  
Cargill Turkey Production, LLC  
George's, Inc.  
George's Farms, Inc.  
Peterson Farms, Inc.  
Simmons Foods, Inc.  
Willow Brook Farms, Inc.

**Prepared by:**

Charles D. Cowan, Ph.D.  
Analytic Focus LLC  
4939 De Zavala Road, Suite 105  
San Antonio, TX 78249

November 26, 2008



---

Charles D. Cowan, Ph.D.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

**PERSONAL SUMMARY**

1. My name is Charles Cowan. I reside in San Antonio, TX. I was retained by the defendants to provide an opinion regarding the use of principal components analysis by Dr. Olsen for this litigation and the statistical reliability and value of sampling used both by Dr. Olsen and Dr. Harwood. I have personal knowledge of the matters contained in this report.

*Education and Experience*

2. My background covers 30 years of research and study in the areas of statistics, economics, and their application to business problems. I am Managing Partner of Analytic Focus LLC, a company headquartered in San Antonio, TX and with offices in Birmingham, Alabama and Washington, DC. A portion of our work is conducting research for legal matters, including providing litigation support and expert witness services when requested. Some of our work focuses on measurement and mitigation of risk for financial intermediaries. The final area of our practice is in support of Federal and State agencies needing economic and financial analysis to pursue their missions. Prior to starting Analytic Focus LLC I served as Chief Statistician for the Federal Deposit Insurance Corporation. I was also a Director for Price Waterhouse where I headed the Financial Services Group in the Quantitative Methods Division. I served for 12 years at the U.S. Bureau of the Census where I was responsible for the evaluation of the Decennial Census and held the title of Chief of the Survey Design Branch.

3. I am currently an adjunct professor in the School of Public Health at the University of Alabama – Birmingham (UAB) and previously served as a professor in the Business School at UAB, as a visiting research professor at the University of Illinois, and in other academic and professional positions.

## REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

4. A listing of my qualifications as an expert in this case are presented in Appendix 1. My complete resume and a listing of all my publications are presented in Appendix 2. A listing of past cases in which I have been deposed or presented testimony at trial is presented in Appendix 3.

*Scope of Assignment & Compensation*

5. I was asked to consider the claims made by the plaintiffs in the above referenced case and to offer an opinion on issues pertaining to their claims. This report considers both issues.

Personnel	Fees per Hour
Charles Cowan, Ph.D.	\$425
Senior Financial Analyst	\$395
Senior Research Associate	\$295
Programmer	\$225
Research Analyst	\$125

For expert representation, depositions and testimony, our hourly rate is \$525. Out-of-pocket expenses, including travel, are billed separately and are in addition to the hourly fees.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

## INTRODUCTION

6. I was asked to review the mathematical and statistical foundations for the use of Principal Components Analysis (PCA) in the report by Dr. Olsen and the selection and use of samples by both Dr. Olsen and Dr. Harwood in their reports. In the former case – the PCA – I looked at what a PCA is, how it was used, what methods were employed in actually performing the PCA, and issues in the construction of the data used in the PCA.

7. In the latter case, the sampling, I looked at how the sampling approach was constructed, and the use of samples in drawing inferences. I examine the ability of Dr. Olsen to draw inferences about the sources of constituents of the watershed and the ability of Dr. Harwood to draw inferences about the characteristics of various animal populations from the sample she used.

8. I concentrate first on the report by Dr. Olsen and examine the methodology he employed, and then move to Dr. Harwood's report. The following sections refer to Dr. Olsen's report and address distinct parts of his work:

I. What is a Principal Components Analysis?

II. What Did Dr. Olsen Do?

A. Collection of data from different sources

1. Consolidation of data into a single dataset
2. Conversion of the data into a form suitable for analysis
3. Problems finding data
4. Problems summarizing into averages

B. Missing Data

1. How much?

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

2. Dr. Olsen's Substitutions

3. Problems with data from multiple sources

C. Methods for Treating Missing Data

1. Means

2. Structural relationships

3. Increases in the Variability of the Data

4. Biases in Correlations

D. Non-Detects

1. Use of substitution to allow for non-detects

2. Variability in detection levels

E. Use of Logarithms

1. Comparison of Logarithms to Original Data

2. Potential reasons for use of logs

3. How Logarithms change the relationship studied

4. How the transformation changes the correlation

5. How the transformation affects the non-detects

F. The Number of Principal Components and Rotations

1. Choice of Principal Components for Analysis

2. Use and non-use of rotations for comparative purposes

III. What Did Dr. Harwood Do?

A. General Principals of Sampling

1. What to Measure

2. How Precise?

3. Representativeness

4. How the Samples Were Selected

**WHAT IS A PRINCIPAL COMPONENTS ANALYSIS?**

9. PCA is a method used to summarize information. In a research study, a researcher will make multiple observations (collect multiple samples) which contain measures on a number of different factors of interest in the study (variables). A common example is taking measurements of weight, height, girth, body mass index, and similar variables on people.

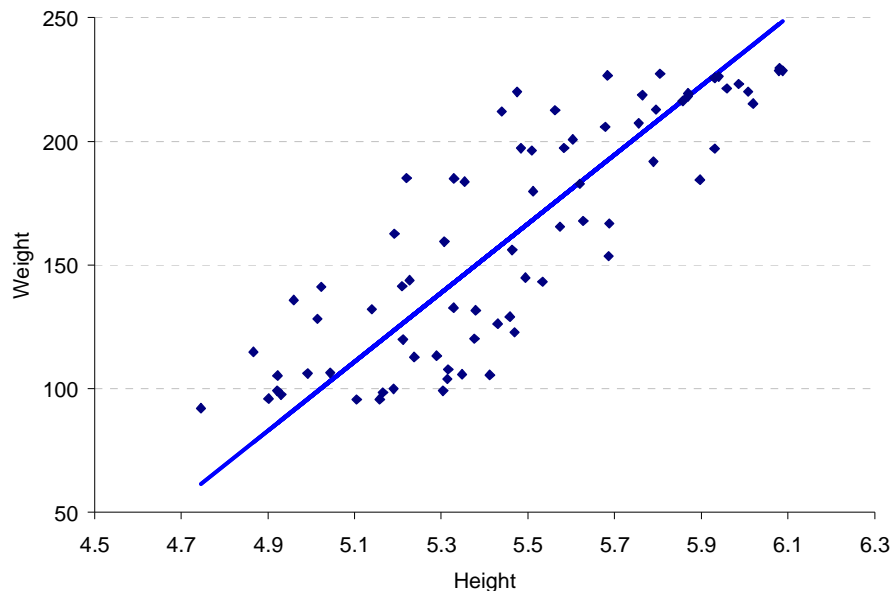
10. The researcher measures multiple variables quantifying characteristics of the sampled item. These variables will be correlated with one another to varying degrees, and so although multiple variables are collected, there will be less real information available to the researcher because of redundancy between the variables. Principal components is a method for examining and summarizing the amount of information actually collected. A simple example follows.

11. Suppose we have the height and weight of a sample of a number of adults. We know these values are related, and if we chart the values we see that there is a distinct pattern to the data. In Chart 1, as height increases, weight increases. However, we could have also turned the chart around and observed that, as weight increases, height increases. The two values are strongly related, but one cannot say that one causes the other – they just increase together. The straight line that runs through the points is the first principal component – it is the line obtained by minimizing the distance from each point to the line, measured at right angles to the line. Not up-down, not left to right, but the shortest distance to the line for each point. This line measures the relationship summarized in both height and weight, although we do not know what this relationship is. We can call it “size” since that seems to be what it is measuring. It is also the line that captures the most variability in both variables. If the variability for height and weight separately are large, it is now summarized in one variable instead of two so that only the new variable (size) has all the variability.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

**Chart 1: Size Measured as Height and Weight**



12. A principal component measures a summary relationship between all variables being analyzed, and does so by describing a line that encapsulates the relationship. The line is written as the sum

of each variable, with a weight on each variable to indicate how much it contributes to the relationship.

13. The line above would be summarized as:

$$(\text{Principal Component 1}) = a_1 * \text{Height} + b_1 * \text{Weight}$$

14. We don't know what this value measures – it is an artificial construct based on the relationship between height and weight. We can imagine an underlying relationship in people called “Size” and that Height and Weight are different manifestations of this value. If we want to summarize the set of measurements in one dimension, we have a single variable we measure called size rather than two variables we measure, like weight and height separately. There is now one dimension instead of two and we have eliminated redundancy in the data. This doesn't seem important in the case of two variables, but with multiple variables, each measuring only a piece of the underlying factor, this can be very important.

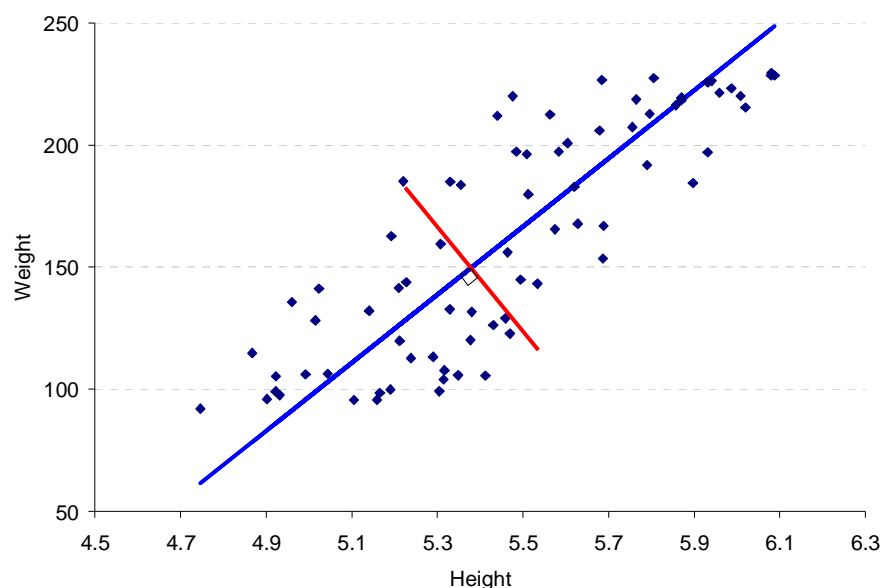
REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

15. The origins of principal components can be found in psychology and education, and are the foundation for IQ tests and educational attainment tests. In education, we don't have a single variable that measures everything we know – it's tested by asking multiple questions and then obtaining a final score on a particular subject. The Scholastic Aptitude Tests (SAT) used as part of the application to get into college are done in this way, testing how much a person knows in a subject. Not all the questions get the same weight – easier questions get less weight than hard ones, and the weights to combine all questions are computed using a technique like Principal Components Analysis.

16. There are as many principal components as there are original variables. The second, third, fourth, and so on principal components are measured at right angles to earlier principal components. Each new principal component captures what is left over of the variability in the data from the previous principal components.

**Chart 2: A Second Principal Component**



17. On Chart 1 a second principal component could be measured at right angles to the first one. This is done on Chart 2. The second component measures a different relationship between height \ weight.



REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

18. The second principal component is a line at right angles to the first. written as

$$\text{Principal Component 2} = a_2 * \text{Height} + b_2 * \text{Weight}$$

19. This second relationship is uncorrelated with the first principal component. It summarizes variance left over. In Chart 2, the line is much shorter than the first principal component because there is less variability left over to explain. Note that the later principal components may or may not measure something of value – they could just be measuring whatever leftover variability there is in the data.

20. Or they could be measuring a unique basis for why the first few principal components do not fully explain the data. Continuing the height and weight example, the further a point is above the first principal component (in blue), the more overweight the person is **relative to the norm defined by the first principal component**. Points below the first principal component are people who are underweight **relative to the norm**. In this example, the main principal component establishes a norm for the relationship of height and weight. The second principal component measures how far one is above or below the norm – much as a physician would decide whether a patient is overweight or underweight.

21. There are four other issues to understand about Principal Components Analysis. These relate to strength of relationship, sampling, interpretation, and utility.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*Strength of Relationship*

22. If the distribution of the (height, weight) points is very close to the line fit through them (the example above was Principal Component 1 =  $a_1 \cdot \text{Height} + b_1 \cdot \text{Weight}$ ) then the relationship is very strong. If the distribution of the points is wide and not close to the line, then the relationship is weak. Each variable contributes “one” to the overall variability. With 26 variables, this means that the overall variability that can be explained is 26.

23. When a solution is found in Principal Components Analysis, each principal component (called an “eigenvector”) has a corresponding measure of how much variability is explained (called an “eigenvalue”)<sup>1</sup>. The eigenvector is the set of weights applied to the variables. In the relationship [ Principal Component 1 =  $a_1 \cdot \text{Height} + b_1 \cdot \text{Weight}$  ], the values ( $a_1, b_1$ ) together are the first eigenvector (a vector is a collection of weights in an equation for a straight line).

24. The eigenvalue ( $\lambda_1$ ) for the first vector ( $a_1, b_1$ ) is a measure of how much overall variability in **all** the variables is explained. The values of ( $a_1, b_1$ ) are chosen so as to make  $\lambda_1$  as large as possible. In other words, the line is the best fit – regardless of how strong or weak the relationship is – to all the points because it explains the most variability. That doesn’t mean it does a great job of explaining the variability. Rather, it’s the best we can do given how strong the relationships are. Chart 3 gives four different relationships from the same example where there is a perfect, a strong, a moderate, and no relationship.

---

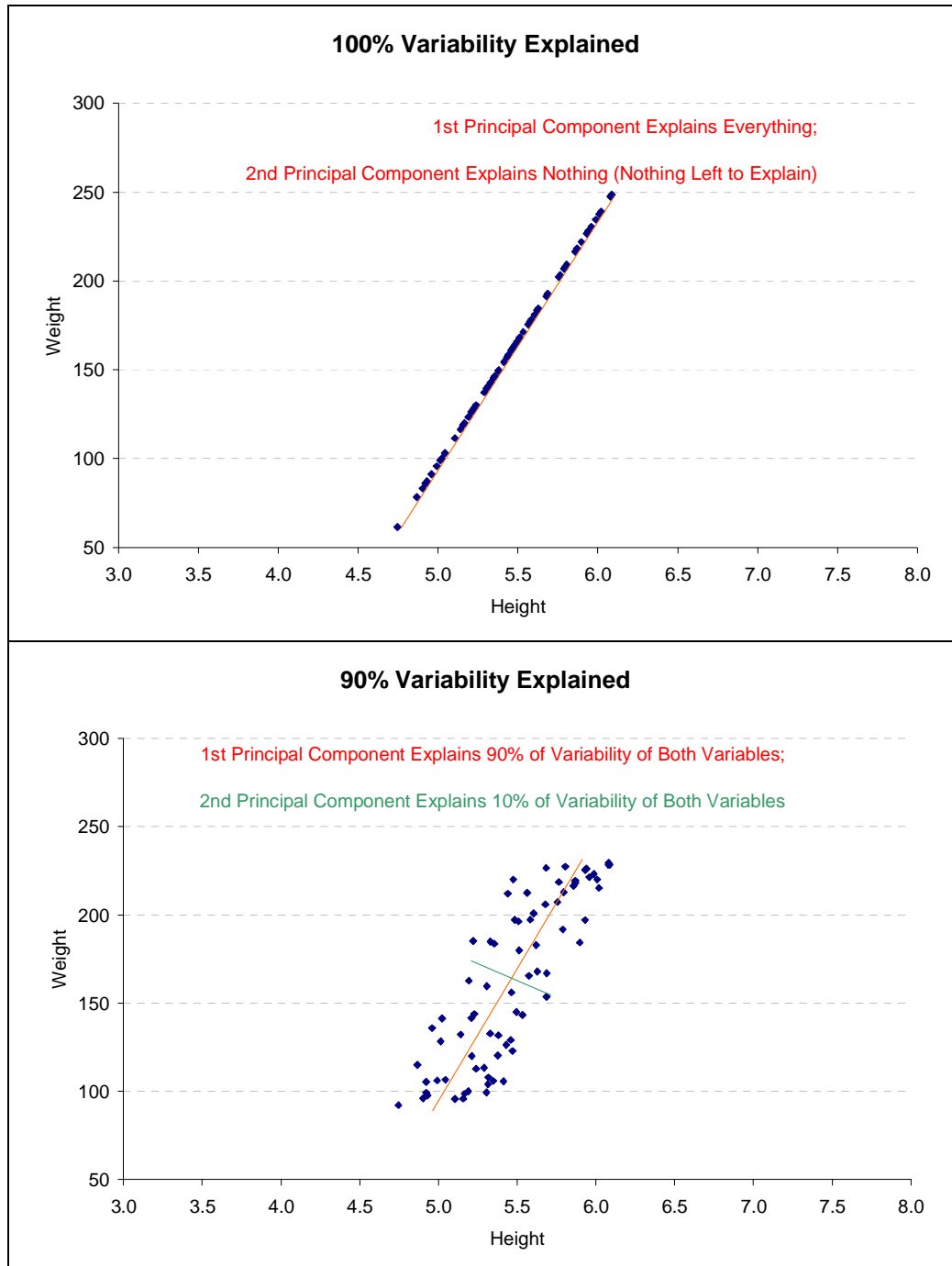
<sup>1</sup> "Eigen" is German, meaning “inborn or forming a natural or inseparable part or quality of”, from Dictionary.com.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

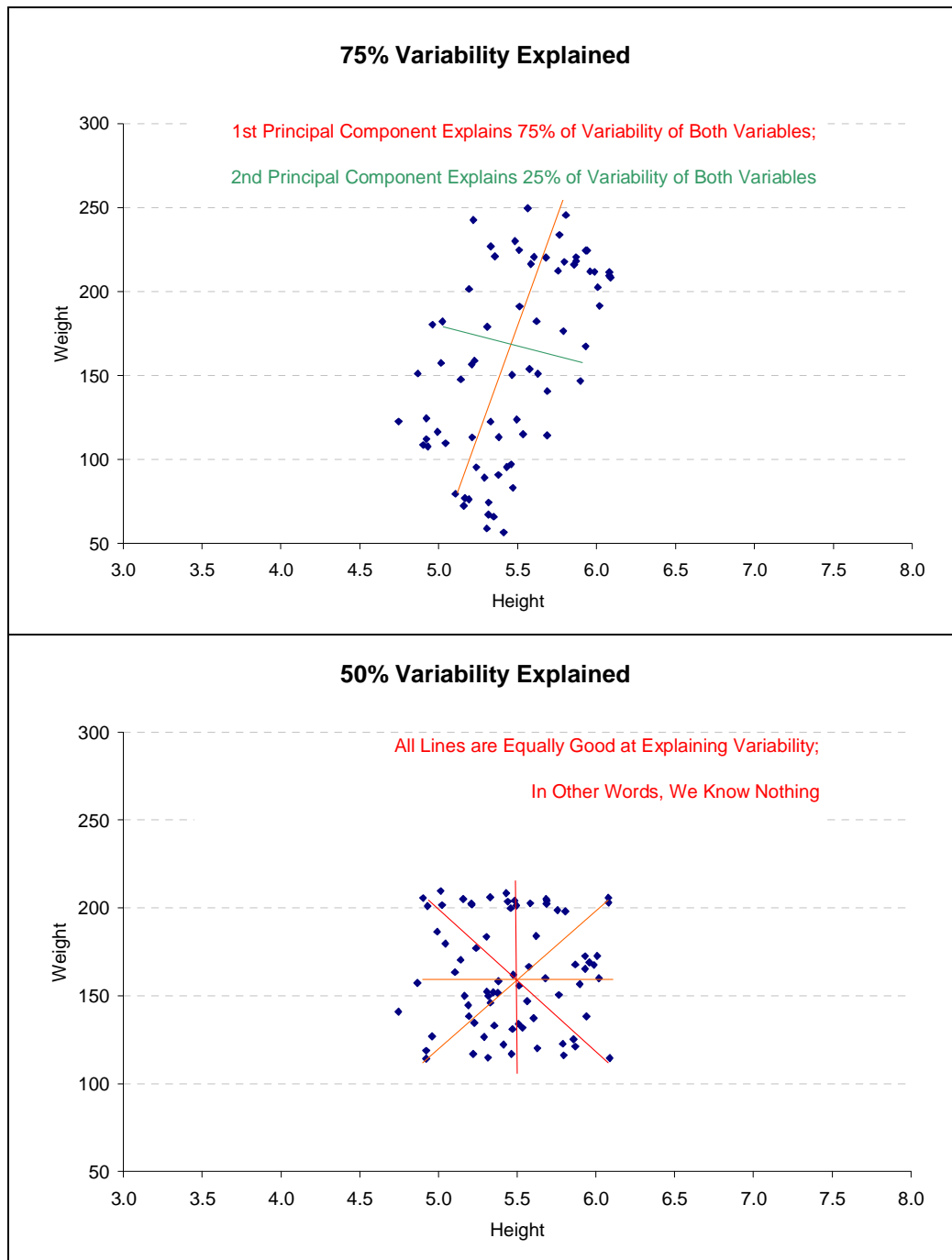
---

### Chart 3: Strength of Relationship Examples

25. The eigenvalues for the next four charts are 2 (out of 2 = 100%), 1.8 (out of 2 = 90%), 1.5 (out of 2 = 75%), and 1 (out of 2 = 50%).



REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED



26. In the bottom chart, the 50% Variability Explained means that Principal Component 1 and Principal Component 2 each explains the same amount – in other words, there is no advantage or new information in the principal components since they don't explain or account for any more variability than the original two variables.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*Sampling*

27. It should be obvious, but it needs to be said: any statistical technique is only as good as the data collected. In particular, if the PCA is based on a sample, then to be able to say something about a population, one needs to have a projectable sample. A projectable sample is one where the methods used to select the sample enable the researcher to extrapolate from the sample to the population. For example, a sample of voters, if selected correctly, can be used to project to the population to forecast the outcome of an election. A group of voters who respond to a CNN on-line poll is NOT a random subset of the population and is meaningless for use in determining what voters in the population are thinking.

28. If Dr. Olsen's sample is not projectable to a broader population or to the area covered in his analysis, then the PCA has no worth in making a statement about what is occurring in the Illinois River Basin. Other reports delve into the quality of the sampling. If it is established that Dr. Olsen's sampling is not representative of the Basin or is biased in some fashion, then the PCA he conducted has no determinative value.

*Interpretation*

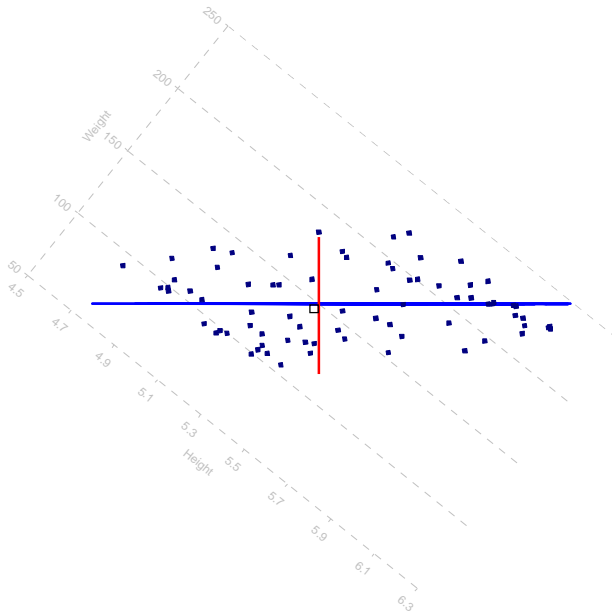
29. When we measure a dimension in a PCA, the question arises as to how to measure the new dimension. In other words, what is a large or small value? In the original data, larger values of height and weight are easy to determine, but that is because they are measured separately on the horizontal and vertical dimensions. The principal component in Chart 1 is hard to interpret as it stands because it requires two dimensions to display it.

30. But the principal component is supposed to be only a single (underlying) dimension. So we can turn the chart so that we can use the line as the dimension we want to summarize.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

### Chart 4: Rotated Principal Component



31. In chart 4 we can now measure “size” on one dimension, and “off-norm” as it’s own dimension going in a different direction. This is the exact same data and the same lines, but rotated to give meaning to distances right and left and up and down

32. The choice of rotated outcomes is important for the interpretation of the data. Dr. Olsen does not consistently

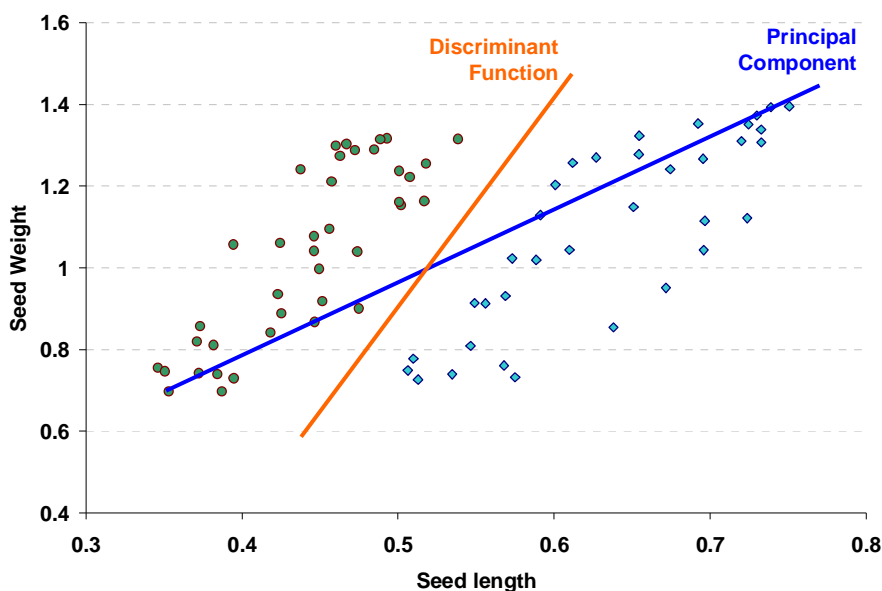
report the rotated values – in fact, he goes from unrotated to rotated solutions without recognizing that there are problems of interpretation. This topic will be discussed later.

### *Utility*

33. Principal components is an excellent technique for discovering a dimension or factor that is continuous and giving values to indicate large or small. It is not usually an appropriate technique for discerning the difference between two groups. Often, a completely different technique should be used to differentiate between two or more groups. In the following example, we have a chart showing weight and seed length for two types of flowers. The principal component can show size, but it can’t be used necessarily to differentiate between the two groups. There are other methods that are much better suited to differentiation. One such technique is discriminant function analysis.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

**Chart 5: Principal Components versus Discriminant Functions**



34. Suppose a grower was trying to differentiate between two types of seeds for sale, as seen in Chart 5. Using a principal component the grower would be wrong half the time, whereas if the grower used a

discriminant function, he could easily distinguish between the two groups.

35. Discriminant Function Analysis (DFA) allows for tests to determine if it is possible to differentiate between two groups; PCA does not. DFA attempts to maximize the difference between two or more groups; PCA does the reverse. PCA homogenizes the data to get an underlying factor. Dr. Olsen chose a technique that requires a subjective judgment regarding how to divide his data into two groups<sup>2</sup>. There are other techniques like the one demonstrated above that give an objective method for determining if it is possible to distinguish between two groups and the test for determining how to best distinguish between the two.

<sup>2</sup> The choice for the threshold is 1.3 for his first principal component. CDM Report, page 6-60

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

## **WHAT DR. OLSEN DID**

36. This section reviews the steps taken by Dr. Olsen to create a data base and conduct his analyses.

### *Collection of data from different sources*

37. For the samples taken by a variety of entities, Dr. Olsen had a data base created made available to us in Microsoft Access<sup>3</sup>. This data combines data collected by the plaintiffs and other samples selected by the USGS<sup>4</sup>. Data is stored as individual observations on specific chemicals or organic matter, identifying the sampling group, the sample within the sampling group, the particular constituent and the amount found in the sample. Different samples had reports on different chemicals or bacteria. There are over 100 indicators measured in the samples, but in each sample there are reports on only some of the 100, so not all samples have measurements on all indicators used in the analysis process.

38. For some samples, there are more than one measurements for a particular indicator (chemical or organic constituents, e.g. bacteria), and so these were averaged in the sample that Dr. Olsen analyzed. Thus, there may be only one observation on aluminum in a sample group, so it is the average. On the other hand, there may be four measurements on fecal coliforms, and the value used from the sample is the average from the four observations<sup>5</sup>.

---

<sup>3</sup> CDM Report, Section 4, Database Compilation and Maintenance, page 4-1

<sup>4</sup> CDM Report, Section 2.10, USGS Sampling, page 2-39 and USGS in DB page 6-38

<sup>5</sup> "EDAnalyzer also has an option for creating (or averaging) the cross-tabulation by sample or by location; e.g., in the case of by location, the data for a particular variable with multiple samples assigned to that location would be averaged during creation of the cross-tabulation", CDM Report, page 6-47

---



REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

39. **This is the first key problem in Dr. Olsen's analysis.** He has samples of different sizes summarized as a single observation for his analysis, and thus all data in the analysis are treated as if it has equal contributions to the variability in the data. The truth is, the variability of the data for the bacteria is much greater than the variability for the remaining 22 variables. The reason that the bacteria are more variable is that there are many more observations for bacteria, and these have additional variability when averaged. He summarizes the bacteria data in one point rather than keeping the distribution of bacteria values. This makes it seem that the bacteria is less variable than it actually is, because the average must be less variable than the original set of multiple observations (a basic principle of statistics). If the real variability of the measured data were represented in the summary database used for the PCA, the relationships between the bacteria and all other values would be greatly different, and the results of the analysis would be greatly different.

40. As it stands, Dr. Olsen does not retain or analyze a principal component that summarizes the bacteria – he throws it away. If the bacteria had the correct variability represented, inclusion of this variability would cause the results to be greatly different. Dr. Olsen did not analyze the variability in the data, though he offers that he has. He has disguised the variability through the averaging process, thus giving too much weight to some variables like phosphorus and not enough weight to other variables like the bacteria.

41. A second outcome results from this flaw – one that is even worse. Since all of Dr. Olsen's principal components are derived from summary measures of variability, the correct calculation of variability would change all of the weights he derived and completely change the outcomes that he presents, and change them in ways we cannot project. This invalidates any of the results Dr. Olsen submits from the Principal Components Analysis.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*Replication of Dr. Olsen's Analysis Dataset*

42. Basically, we can't. Although we followed the paradigm described above on all the Access data sources presented to us (specifically Access database 20), we could not replicate the initial analysis dataset (the Excel subdatabase SW3) exactly. The sequence of construction is to convert the Access database to a large Excel database to a summary extract used for the actual PCA. However, **it is not possible** to go from the Access database, which is the original repository database, to the database used for the PCA analysis. For most observations, we can replicate the outcomes exactly. With some additional guesses as to the treatment of unusual observations, we were able to replicate more. But there are still a number of observations where we cannot exactly replicate the data that Dr. Olsen analyzed. A more complete description of how observations were replicated is given later in the report.

43. **This is the second key problem in Dr. Olsen's analysis.** Dr. Olsen or his colleagues were inconsistent in their treatment of the original observations to obtain the dataset they analyzed. There is random noise or possibly bias introduced in the data he analyzed since he followed different procedures for different observations. Because of this, the data he analyzes represents different things since the different treatments mean that not all measurements are measuring the same thing. Further, a real scientific study should be able to be replicated by another scientist following the procedures of the first. We cannot – there is no way to take a single set of procedures, either as outlined by Dr. Olsen or modified through detective work, to obtain the dataset that Dr. Olsen analyzed.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*Missing Data*

44. As noted above, not all samples have measurements on all observations. In fact, there is a very significant amount of missing data. Dr. Olsen disguises this by substituting for the missing data. He plugs in the mean of a variable for the actual (though missing) value. Only 267 of the 573 samples used by Dr. Olsen have complete data. This means only 47% – less than half – of the observations have real data actually observed in the field. This means that more than half of Dr. Olsen's observations have data that Dr. Olsen substituted rather than real data.

45. **This is the third key problem in Dr. Olsen's analysis.** Dr. Olsen has plugged in so many missing values that a very significant part of the dataset is **made up** by Dr. Olsen. While he analyzes both the data set with no records with missing data and a second data set with substituted data, he fails to admit that he has plugged in values that skew the correlational structure. Dr. Olsen substitutes the mean for a missing value<sup>6</sup>: if aluminum is missing, he substitutes the mean for aluminum from the other sampling sites where aluminum was recorded. This means that he can take data from sites that are in his view poultry impacted and substitute this data into a site that he would not consider being poultry impacted, completely skewing the dataset to show what he wants to show.

46. Dr. Olsen was also missing data in a second way. Samples selected by the USGS measured some different values in the chemicals or processed and tested them in different ways. In particular, total dissolved solids, Sulfate, Total Kjeldahl Nitrogen (TKN) and all three measures of phosphorus were all analyzed in a different way<sup>7</sup>. There are other differences

---

<sup>6</sup> There is no direct statement in the CDM Report that states missing values are replaced with means but replacing missing values with means is the only way to reproduce the results from Dr. Olsen's analysis.

<sup>7</sup> CDM Report, page 6-36

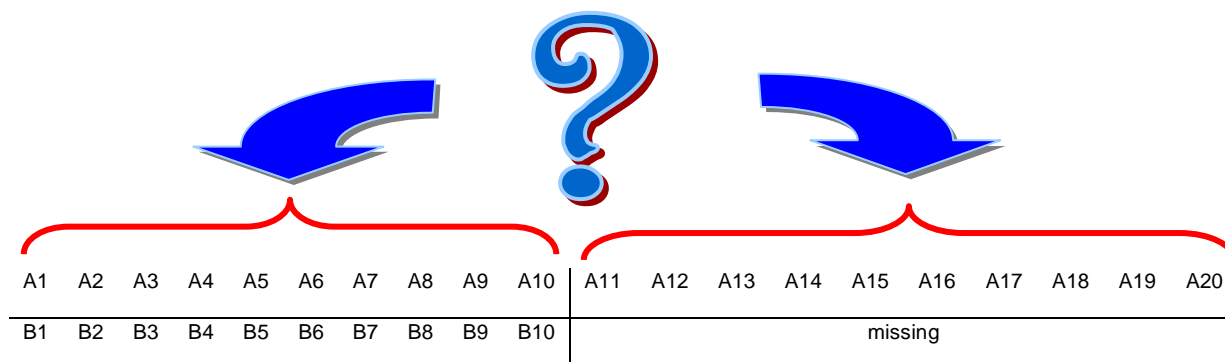
REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

between the USGS data and the remaining data, many of which may be even more substantial in their impact (for example, flow rates differed in the two sets of data). Dr. Olsen doesn't conduct any test for the implication this might have on the PCA. Such a test is presented later in this report.

**47. This is the fourth key problem in Dr. Olsen's analysis.** Dr. Olsen used data from two sources as if they were equivalent, without testing to see if they measured the same outcomes. This is completely contrary to scientific method, and in particular is a procedure that undercuts the analysis Dr. Olsen is trying to perform since it is another source of variability in the data.

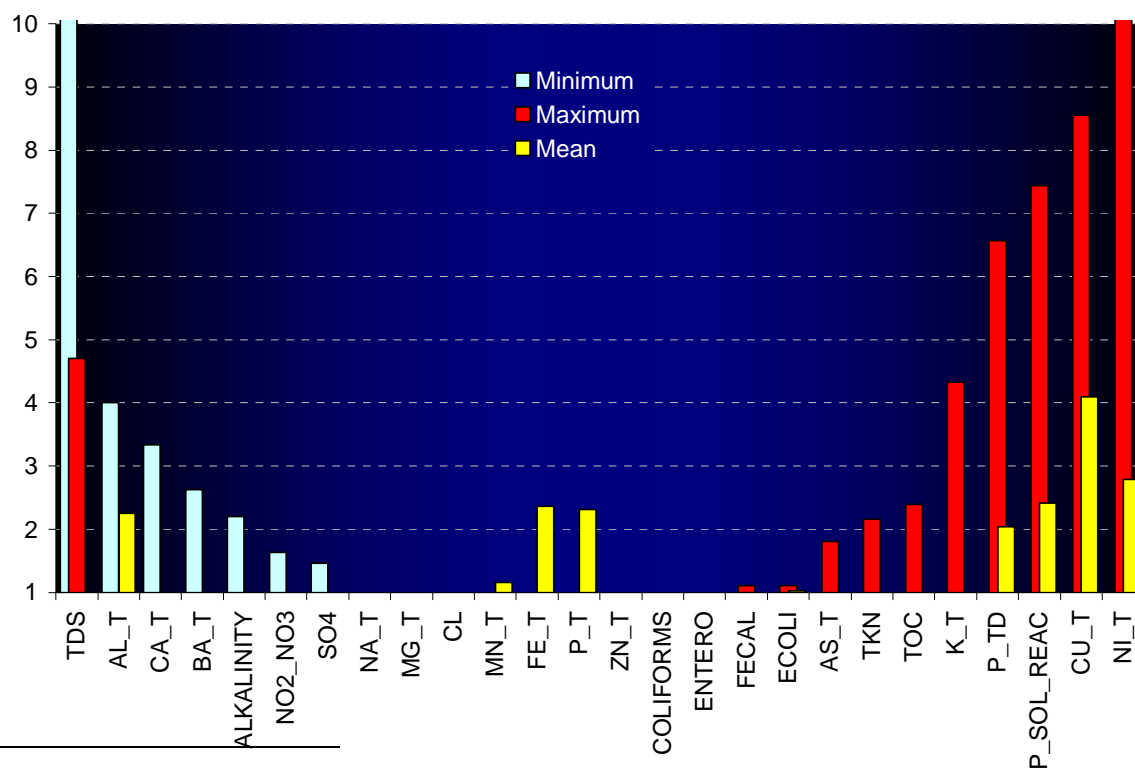
48. Finally, in looking at observations with and without missing data, if the data were just missing at random, we would expect that values in cases with missing data would be just like values in cases without missing data. Suppose we have only two variables: A, and B. Variable A has all of its observations, variable B is missing half of its values. We divide the data into two sets: observations that have measurements on both A and B, and observations that have measurements on only A. If the measurements on B are missing at random, then we would expect values in the first half of A to be like observations in the second half of A.



REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

49. When we examine the data from Dr. Olsen's files, we find this isn't even remotely true for the 26 measures he uses<sup>8</sup>. We looked at the means from the 267 observations that had no missing data. We compared these to the means from the 306 observations that had some missing data. Because the 26 different variables have different scales, we took the ratio of the mean from the first group (no missing) to the mean of the second group (some missing). If the means were the same, this ratio would be unity (1.0). If the ratio was between zero and one, we inverted the value so that it would be measured on the same scale from one to infinity. We performed the same operations for the minimum values of these two sets and the maximum values for these two sets for each variable. All three ratios are expected to be equal to one if the two groups (Group 1 = not missing versus Group 2 = missing) are the same.

**Chart 6: Ratios of Mean, Minimum, and Maximum for Observations with Some Missing versus Non-missing**



<sup>8</sup>CDM Report, page 6-45

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

50. There are only eight variables out of the 26 where there isn't some serious difference between the two groups (missing and non-missing). Variables on the right side of the chart have significant differences in the maximum values, meaning the range of values for one group is truncated relative to the other. Variables on the left side of the chart have significant differences in the minimum values, meaning again that the range of values for one group is truncated relative to the other.

51. **This is the fifth key problem in Dr. Olsen's analysis.** These inconsistencies mean that the data is severely biased by missing values. Observations that are missing some data are unlike those that are not missing data. Analysis of a data set with characteristics like this is fruitless since there is no way to know what the real relationships are in the data. Since PCA relies on these relationships, the PCA conducted by Dr. Olsen is meaningless.

*Methods for Treating Missing Data*

52. Dr. Olsen substitutes the means for the missing data. This forces the distribution of the data to change since different variables have different numbers of missing values. It would seem that this process would reduce the variability in the data set, but in fact it may artificially reduce the variability on an individual variable. At the same time it will also inflate or deflate the correlation between two variables, and change the direction of the correlation.

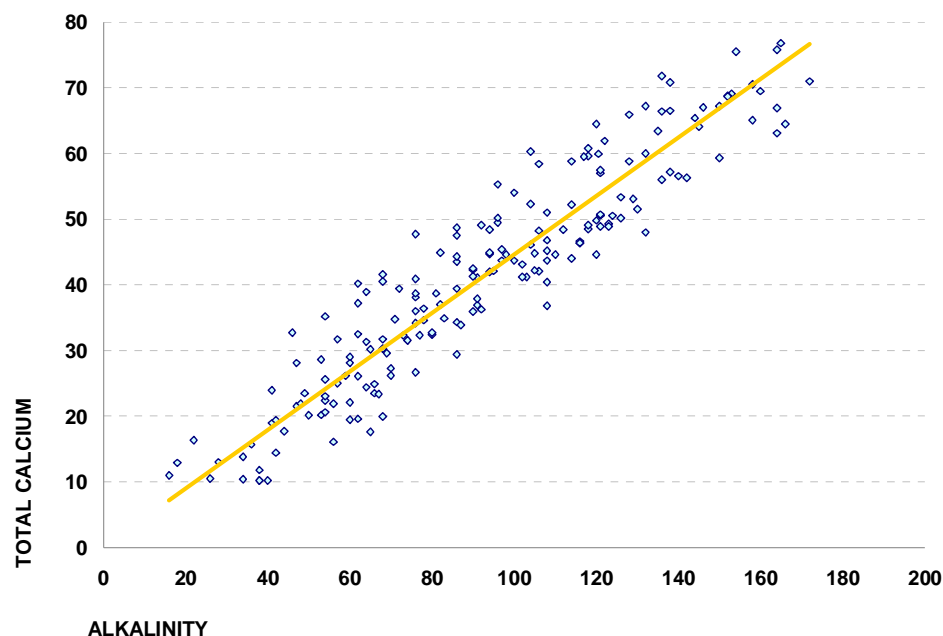
53. An example taken from Dr. Olsen's data follows. Chart 7a shows the relationship between calcium and alkalinity for observations where both values are observed.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

**Chart 7a:**

**Complete**

**Observations**



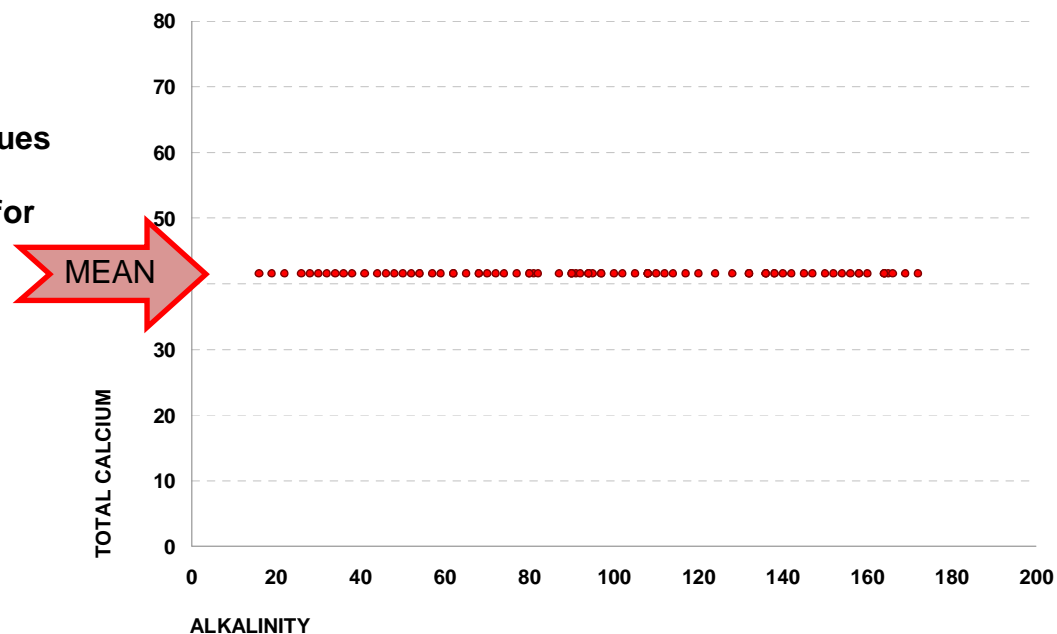
54. Dr. Olsen is missing a large number of observations on both Calcium and Alkalinity. When he is missing an observation, he substitutes the mean, regardless of what he knows about the other variable. In other words, if he is missing a value on Calcium, he plugs in the mean regardless of anything he knows about alkalinity. The same is true for alkalinity.

**Chart 7b:**

**Missing Values**

**Plugged In for**

**Calcium**



REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Chart 7c:

Missing

Values

Plugged

In for

Alkalinity

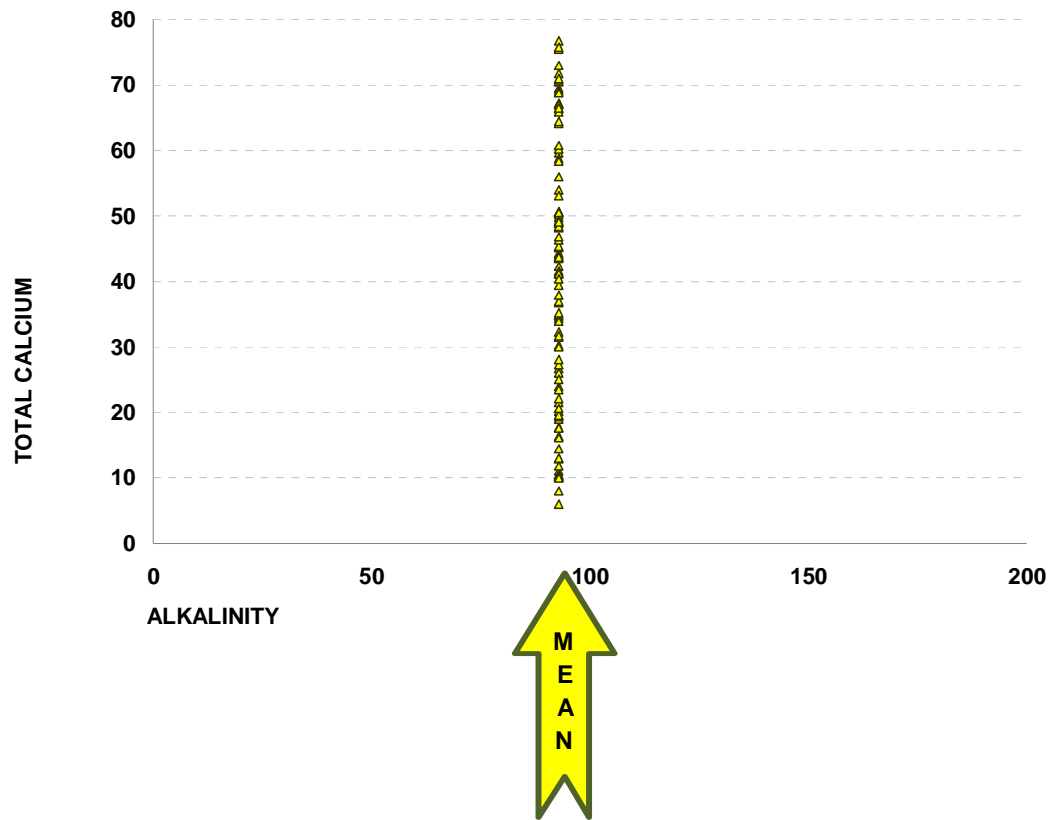
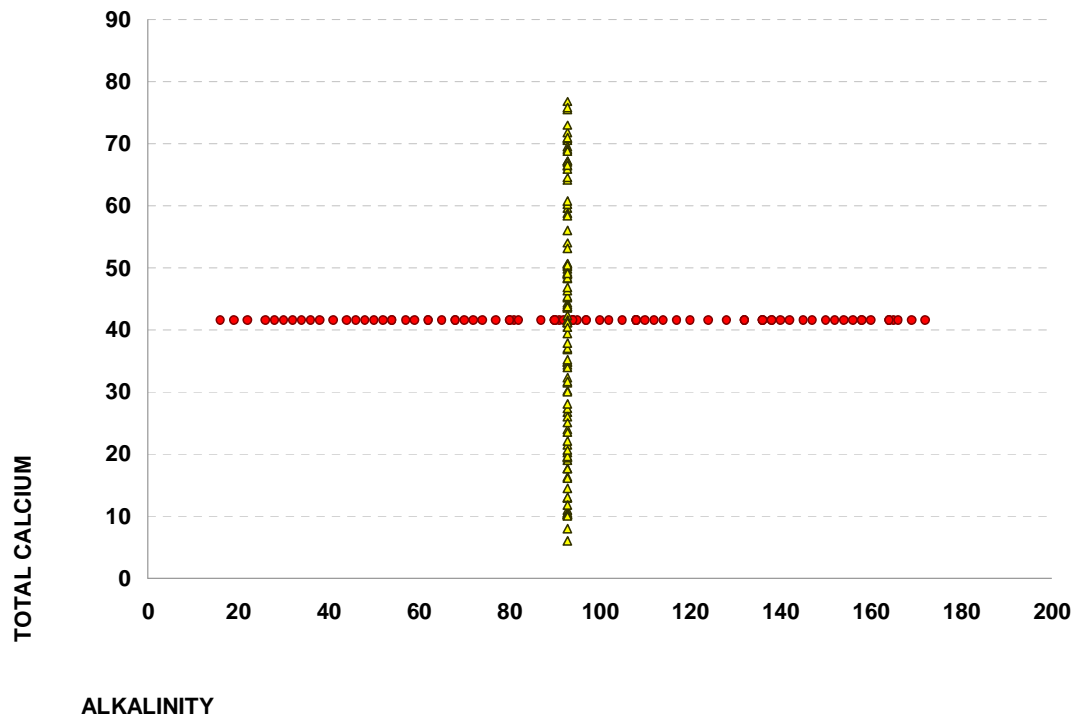


Chart 7d:

Combination

of Missing

Values

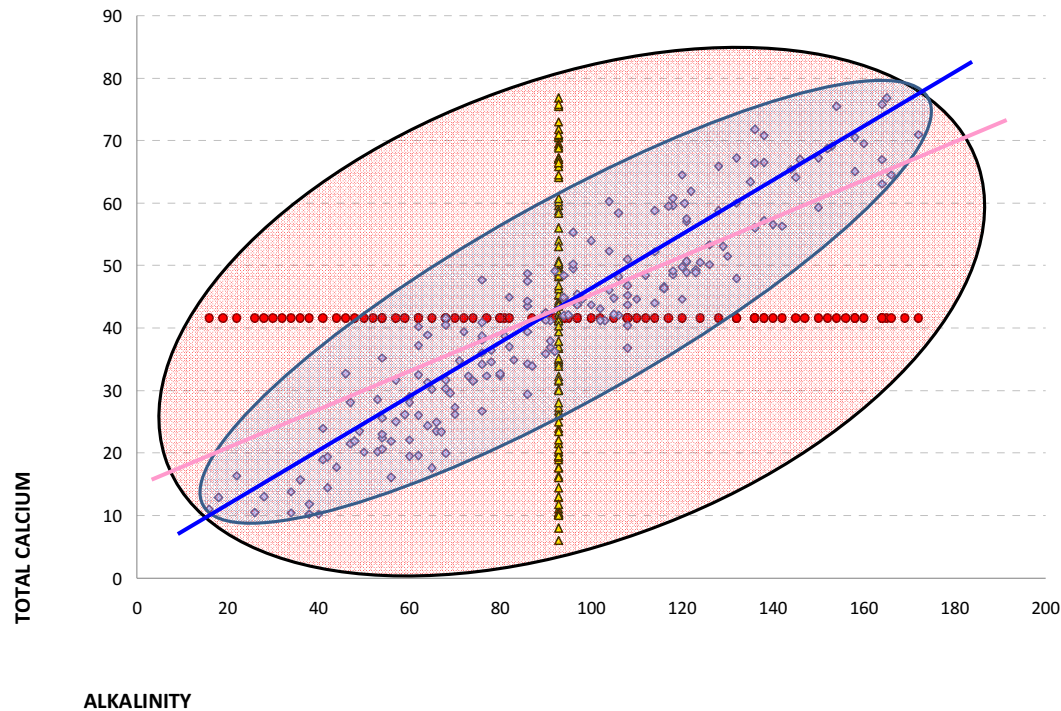




REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

**Chart 7e: Combination of Missing Values with Known Values – the Data Set**

**Analyzed by Dr. Olsen**



*The larger pink ellipse covers what Dr. Olsen analyzed, but it is skewed from the real data and has a much greater artificial variability. The narrower blue ellipse is the original data.*

55. Any line used to describe a relationship is skewed by the amount of missing data substituted and the differences in the ranges and means of each set of data (missing and nonmissing).

56. **This is the sixth key problem in Dr. Olsen's analysis.** His method of substituting for missing data skews the relationships in the data. At the same time, Dr. Olsen's method of substitution inflates variances, again changing the relationships being measured. These two outcomes make it impossible to measure any true relationships in the data. Dr. Olsen has hidden the true relationships by changing them with missing data substitutions designed to hide defects in his data and his calculations.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*Non-Detects*

57. In the data analyzed by Dr. Olsen, he also has a number of values that are non-detects, meaning the measurement method used by the researchers cannot measure any trace measure of a chemical or organic value. Rather than treat this as a zero (not detected), Dr. Olsen substitutes the midpoint between zero and the detect limit for a chemical<sup>9</sup>. However, the detect limits can vary from observation to observation for each chemical. In some samples we would have a smaller non-detect than for others, such as .01 as a lower limit for some observations on Aluminum, and .001 for other lower limits. This variability in detection levels adds to the variability in the data, exacerbated by the use of logarithms. This is another method of treatment of missing data, but the impact will be discussed later in this report.

*USGS vs. non-USGS observations*

58. As noted previously, Dr. Olsen takes observations from the USGS<sup>10</sup> and combines them with observations from the plaintiffs and treats them all as if they are measuring the same relationships, but he does so without testing if there is a difference between the two datasets.

59. **This is the seventh key problem in Dr. Olsen's analysis.** Ignoring the sources of the data ignores any incompatibility in the data. The table below replicates Dr. Olsen's analysis exactly for the PCA, but conducts his analysis twice – once for the USGS cases and once for the non-USGS cases. The rotated factors are presented.

---

<sup>9</sup> CDM Report page 6-40 and page 6-47

<sup>10</sup> CDM Report, page 5-1 and page 6-38

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

**Table 1: Analysis of Two Separate Parts of the Data Collected**

NOT USGS						USGS				
Variables	1	2	3	4	5	1	2	3	4	5
MN	0.836	0.102	-0.035	-0.148	0.068	0.897	-0.079	-0.074	0.204	0.046
FE	0.853	0.205	-0.187	0.124	-0.169	0.864	-0.172	-0.013	0.210	0.267
AL	0.787	0.217	-0.263	0.170	-0.209	0.846	-0.162	0.072	0.208	0.284
NI	0.762	0.135	0.224	0.236	-0.049	0.774	0.343	0.103	0.150	-0.222
AS	0.745	0.038	0.106	-0.018	0.013	0.767	0.133	0.303	0.190	-0.342
BA	0.590	-0.032	-0.333	0.016	0.314	0.701	0.320	0.251	0.142	-0.307
CU	0.698	0.308	0.152	0.337	-0.175	0.784	-0.057	0.129	0.085	-0.124
ZN	0.688	0.081	0.093	0.271	0.033	0.881	-0.069	0.070	0.111	0.013
TOC	0.607	0.439	0.306	0.167	-0.223	0.726	0.033	0.028	0.379	0.044
P_SOL_REAC	0.253	0.061	0.318	0.814	-0.079	0.056	0.388	0.861	0.125	0.038
P_TD	0.304	0.089	0.377	0.786	-0.119	0.339	0.307	0.832	0.166	-0.148
P	0.558	0.141	0.283	0.684	-0.126	0.577	0.185	0.702	0.242	-0.139
NO2_NO3	-0.144	-0.039	-0.086	0.734	0.233	-0.229	0.202	0.704	-0.207	0.432
FECAL	0.095	0.954	-0.014	0.048	-0.062	0.316	-0.101	0.097	0.848	0.145
COLIFORMS	0.147	0.913	-0.013	0.039	-0.086	0.312	-0.131	0.180	0.603	0.410
ENTERO	0.130	0.886	-0.035	0.033	-0.090	0.398	-0.195	0.069	0.753	0.273
ECOLI	0.121	0.814	-0.032	0.018	0.026	0.186	0.008	-0.010	0.874	-0.101
CA	-0.133	-0.168	0.183	0.000	0.882	-0.272	0.815	-0.120	-0.107	-0.100
ALKALINITY	-0.100	-0.077	0.216	-0.093	0.835	-0.343	0.626	-0.269	-0.067	-0.075
TDS	0.282	0.003	0.219	0.324	0.476	-0.104	0.871	0.176	0.007	-0.154
SO4	0.109	0.018	0.802	0.110	0.113	0.076	0.864	0.367	-0.083	0.088
NA	-0.223	-0.087	0.837	0.190	0.211	0.063	0.914	0.309	-0.050	-0.004
CL	-0.154	-0.062	0.753	0.217	0.310	-0.021	0.910	0.285	-0.077	0.008
MG	0.492	0.091	0.618	0.077	0.153	0.317	0.757	0.150	-0.077	0.069
K	0.519	0.139	0.537	0.466	-0.097	0.454	0.748	0.412	0.050	-0.046
TKN	0.271	0.322	0.220	0.013	-0.122	0.072	0.028	-0.008	-0.369	-0.717

*Variables in yellow are the six variables that were measured differently by the USGS and the plaintiffs.*

60. Results change significantly for variables that differ in measurement between USGS and the plaintiffs' collection. Note that for the three phosphorus measures, they are only in the fourth principal component for the non-USGS data, but in the third principal component for the USGS data. This means the supposed importance of the phosphorus measures is lower for one data set than another – this shouldn't happen if the two datasets are equivalent.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

61. For the plaintiffs' measurement of total dissolved solids (TDS), TDS doesn't surface on ANY of the principal components, meaning it is not important to any of the factors measured. However, for the USGS measurement, it is a key component of the second principal component. Similarly, in the measurements by the plaintiffs, calcium and alkalinity contribute a completely separate factor, uncorrelated with a principal component that includes sulfate, sodium, chlorine, and magnesium. For the USGS, there is a single component that combines calcium, alkalinity, total dissolved solids with the factor measured by the plaintiffs. Again, this shouldn't happen if the two sets of data are equivalent in the way they measure constituent elements. Finally, TKN doesn't carry any weight in defining principal components in the data from the plaintiffs, whereas in the USGS data it is so important that defines it's own principal component, again separate from the remainder of the components.

*Use of Logarithms*

62. Dr. Olsen converts all of his observations by taking logarithms of values before conducting the PCA<sup>11</sup>. Use of logarithms in statistical analysis is common in a number of fields and is usually done for one of three reasons. One reason is to stabilize the variability of the data so that the data more closely follows a particular statistical distribution. As Dr. Olsen didn't conduct any statistical tests, this can't be the reason.

63. The second reason is to transform data with exponential relationships to data with linear relationships, as methods for analyzing linear data are much easier to employ. This may be the case here, but there are costs for doing so and there is no discussion regarding why data in the water samples would have multiple exponential relationships.

---

<sup>11</sup> CDM Report, page 6-46

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

64. The third reason is to reduce the natural variability of data and pull in outlying observations. When this is done, it typically disguises problems in data collection or unusual observations that should have been separately analyzed. This is certainly the case with this data.

65. The problem with the use of logarithms is that it significantly reduces the variability of the observed data and changes the correlation between the observations. In the extreme, two variables that have no linear relationship (correlation of zero) can have a perfect correlation when one takes logarithms. This means that a PCA of data where logarithms are taken will result in a completely different outcome than a PCA of the original data.

66. Dr. Olsen doesn't explain why he takes logarithms, he simply does so. There is no examination of whether correlations measured on the logarithmic scale also exist in the real world. Dr. Olsen doesn't consider the interpretation of a principal component once he has conducted an analysis.

67. As described in an earlier section, each principal component is a weighted sum of the variables in the analysis. A principal component is written as:

$$\text{Principal Component} = c_1V_1 + c_2V_2 + c_3V_3 + \cdots + c_{26}V_{26}$$

where the coefficients  $c_j$  are related to those presented in the table above in the USGS \ non-USGS analysis or any of the other PCA analysis. However, this would be true for those cases where the variables  $V_j$  are in their original form. Now suppose we have a principal component that is on the logged values.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

68. In Dr. Olsen's analysis, we have:

$$\text{Principal Component} = c_1 \text{Log}(V_1) + c_2 \text{Log}(V_2) + c_3 \text{Log}(V_3) + \dots + c_{26} \text{Log}(V_{26})$$

69. Using a simple algebraic result, we transform this equation into one involving the original data

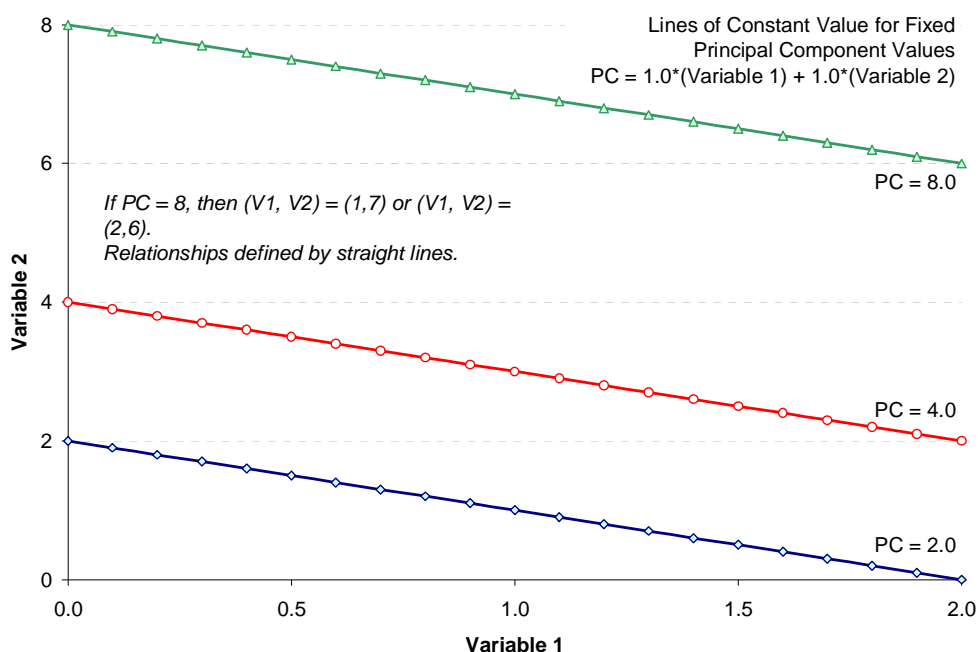
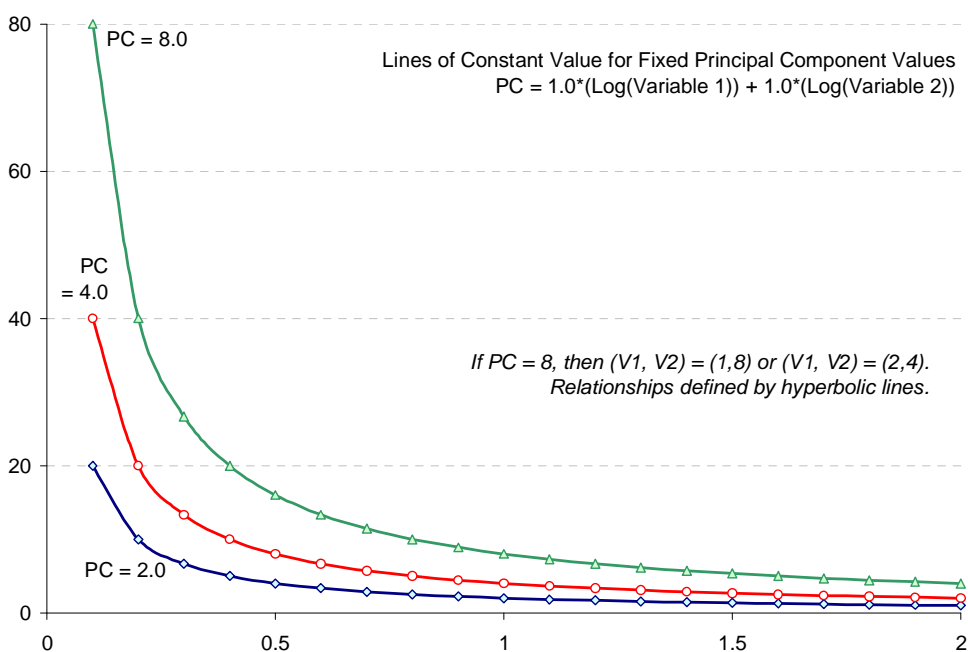
$$\begin{aligned} \text{Principal Component} &= c_1 \text{Log}(V_1) + c_2 \text{Log}(V_2) + c_3 \text{Log}(V_3) + \dots + c_{26} \text{Log}(V_{26}) \\ &= \text{Log} \left[ (V_1)^{c_1} * (V_2)^{c_2} * (V_3)^{c_3} \dots (V_{26})^{c_{26}} \right] \end{aligned}$$

70. With the logarithms, the principal component is NOT a sum of the variables. The principal component is the **product** of the variables, each raised to some factor that weights it. Because it is a product, this means that any findings do not relate back to any findings in the real world in the ways that Dr. Olsen describes. Results from Dr. Olsen's analysis are multiplicative, not additive. Dr. Olsen mistakenly ignores this outcome in his transformations.

71. Charts 8a and 8b show the contrast in the relationships. If Dr. Olsen had not used the logarithms, his relationships between variables would be straight lines. A principal component would represent the sum of values (like a measure of iron plus a measure of aluminum plus a measure of copper within one sample). But Dr. Olsen did use logarithms, which forces all the relationships to be curved. Worse, for a set value in the principal component analysis, the outcome is either a very large amount of variable one combined with a very small amount of variable two (lots of iron, very little copper) or a large amount of variable two combined with a very small amount variable one (very little iron and lots of copper). For the most part, using logarithms, a fixed value for a principal component represents extremes of one variable or another, but not of both variables, completely undercutting his argument that his results represent a "signature".

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

**Chart 8a: Linear Relationships From Use of Actual Variables in Principal Components****Chart 8b: Curved Relationships Implied by Logarithmic Transforms of Variables Used in Principal Components**

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

72. There are other problems with the use of logarithms. One is that, in trying to fit a relationship between two variables, observations receive different weight for their contribution to the relationship if log values are used compared to when the original values are used. This means that there are values that will have a strong effect on the outcomes when used as an actual value. The same values will not have an effect on the outcomes if logged, while other observations will have a stronger effect than would have happened with the original data. Because of this, use of logarithms has to be done with great caution since the interpretation of the value of the inputs differs greatly. A particular example of this is found in the non-detects.

73. As noted before, the **non-detects** have their importance greatly heightened in the analysis. The logarithm of a number is the exponent of the number represented as raised to the power of ten. The table below demonstrates what the values are:

Number	0.000001	0.0001	0.01	1	100	10000	1000000
Equals	$10^{-6}$	$10^{-4}$	$10^{-2}$	$10^0$	$10^2$	$10^4$	$10^6$
Logarithm	-6	-4	-2	0	2	4	6

74. A non-detect of .01 versus a non-detect of .001 might not seem like much a difference, but in the log scale this can be the difference between -2 and -3. If the variable being measured typically has values in the range of 10 to 100 milliliters, the value being analyzed on the log scale is somewhere in the range of 1 to 2. A change in the non-detect value of -2 to -3 (merely because of very minor differences in the test) will have huge effects on the outcome.



75. **This is the eighth key problem in Dr. Olsen's analysis.** He doesn't perform any sensitivity analyses to determine if the non-detect limits affect his outcomes. If most of the values for a logged variable range from 1 to 2 and then an arbitrary value of -2 or -3 is thrown into the analysis, he has created outliers that leverage the relationship. Two variables with a straight line relationship, both measured on a scale of 1 to 2, will be greatly impacted by values thrown in at the far end of the scale. Furthermore, why chose the midpoint between zero and the non-detect value as the substitute value? Why not another value, closer to zero or closer to the non-detect value? Since the log transformation has such power in moving the end of the relationship, the impact of this choice should also have been measured.

#### *The Number of Principal Components and Rotations*

76. The final analytical issue for discussion is the number of principal components that came out of the analyses and their meaning. Dr. Olsen conducted the PCAs as described above, but he only retains the first two principal components. He throws away significant results that may explain patterns not found in the first two components<sup>12</sup>. These later components are the ones that may be most useful in explaining specific results.

77. Further, he arbitrarily reports on non-rotated factors at times and ignores the rotated outcomes. The problem with doing this is that a non-rotated factor is measuring a distance in a way that cannot be interpreted (see the earlier description of this problem). Dr. Olsen's data show the problems with both of these actions. The following table presents the outputs from Systat (the program he used) using the data from his datasheets.

---

<sup>12</sup> "These variances indicate that PC1 and PC2 are by far the most important of the five together explaining 56.2% of the total variance, relative to PCs 3, 4, and 5 (17.8%)", CDM Report, page 6-51

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

**Table 2: The Five Principal Components from Dr. Olsen's Analysis**

CU_T	0.851	-0.032	-0.077	-0.070	-0.047	0.161
P_T	0.812	0.341	-0.057	-0.313	0.142	0.033
TOC	0.812	-0.040	0.110	0.005	-0.175	-0.044
NI_T	0.801	0.106	-0.216	0.078	-0.089	-0.073
FE_T	0.797	-0.332	-0.308	0.021	0.083	-0.203
AL_T	0.765	-0.367	-0.277	-0.043	0.129	-0.195
K_T	0.743	0.473	0.010	-0.135	-0.138	0.020
ZN_T	0.721	-0.075	-0.175	0.130	-0.021	0.248
AS_T	0.672	-0.063	-0.304	0.214	-0.135	0.118
MN_T	0.658	-0.206	-0.385	0.304	-0.029	-0.250
P_TD	0.637	0.511	0.072	-0.423	0.162	0.099
MG_T	0.575	0.422	-0.015	0.259	-0.257	-0.025
P_SOL_REAC	0.559	0.526	0.076	-0.438	0.240	0.115
NA_T	-0.003	0.838	0.259	0.064	-0.223	-0.207
CL	0.036	0.816	0.231	0.132	-0.142	-0.157
SO4	0.243	0.696	0.177	0.102	-0.298	-0.313
TDS	0.302	0.474	-0.092	0.247	0.269	0.131
FECAL	0.554	-0.380	0.651	0.170	0.118	-0.037
COLIFORMS	0.556	-0.370	0.603	0.134	0.103	-0.041
ENTERO	0.552	-0.406	0.578	0.148	0.116	-0.093
ECOLI	0.481	-0.321	0.547	0.231	0.106	0.067
BA_T	0.381	-0.108	-0.460	0.287	0.292	-0.139
CA_T	-0.252	0.538	-0.053	0.609	0.351	0.070
ALKALINITY	-0.227	0.478	0.000	0.649	0.240	0.206
NO2_NO3	0.044	0.406	0.070	-0.320	0.578	0.041
TKN	0.347	-0.044	0.020	0.092	-0.355	0.689

\* Factor loadings above 0.6 are in red, factor loadings above 0.45 are in blue if there are no other factors in red for the same variable.

78. Using Dr. Olsen's methods, we would throw away the principal component that has bacteria (fecal, coliforms, entero, and e-coli). But the other experts for the plaintiffs claim this to be the most important data for analysis of chicken waste. This inconsistent treatment of key information raises the question about what is significant and whether there is any consistent treatment of the data produced by Dr. Olsen.

**REPRODUCING THE SW3 DATA RECORDS AND VALUES**

79. Dr. Olsen used a program called EDA\_Analyzer to capture the data from the main database and loaded the data into an Excel worksheet referred to as SW3. It appears that he substitutes means for the missing values (see the earlier discussion in this report on this point). Dr. Olsen then takes logarithms of the SW3 values before using Systat to calculate PCA loadings (coefficients). The results of the Systat loading coefficients are transferred to an Excel sheet and he calculates the PCA values presented in Appendix F of the CDM report<sup>13</sup>.

80. We attempted to reproduce the values in the SW3 Excel sheet and the PCA values in Appendix F of the CDM report. All of the records from the master database with “SW:S” in the sample groups were downloaded into an Excel file. This download produced an Excel sheet with all of the surface water data. We had to make some changes in sample group identifications to match the EDA\_Sample IDs found in Appendix F of the CDM report. Some of the changes were to add USGS to the sample group IDs that only had numbers. There were other changes made to the sample group IDs that involved removing blank spaces and changing noncapital letters to capital letters. This work was required to be able to finally link the values reported by Dr. Olsen in his written report to the same values in Dr. Olsen’s data – there was little correspondence between values in the written report and the database and it required a significant effort to be able to link which data records Dr. Olsen selected from all of those available. I revisit this topic later as there seems to be little consistency in choices made for the data ultimately included in the analysis. We picked the appropriate measurement unit for values that were measured in UG/L units and we used the P0065 measurement values for the USGS variables TKN, TDS, SO4, P\_TD, P\_T and P\_SOL\_REAC.

---

<sup>13</sup> CDM Report, page 6-53

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

81. In the CDM report, the names of the variables are used in Dr. Olsen's descriptions. The database also has the names of the variables, but a "ParamID number" is also associated with each variables. If the CDM report in Dr. Olsen's tables presented the ParamID number along with the name of the variable, then it would be clear which variable is being discussed. This is an issue because it is nearly impossible to scale down from the 315 variables in the Access database to the 26 in the Excel database. There is no documentation as to exactly which variables were extracted by Dr. Olsen or his subordinates, and only through diligent detective work was it possible to work backwards to discover which 26 variables were selected. As will be seen in a later section, there is no standard data selection procedure that would indicate how Dr. Olsen got from the 315 variables in his full database to the final 26 variables he selected. In fact, given how much information is missing in the database, the final set of 26 variables used is counterintuitive.

82. There are 26 variables in the final SW3 Excel spreadsheet analyzed by Dr. Olsen<sup>14</sup>. Each variable has a parameter key in the database table RefParm that indicates the name of the variable. Appendix F in the CDM report has a listing of the 573 samples (EDA\_Sample) used in the PCA runs for the SW3 data. The EDA\_Sample IDs are produced from combining several sample keys and sample groups into one sample group, or as referred to in the CDM report, an EDA\_Sample. Below is a small example of what occurs when the data is downloaded from the database. There are usually several samples for each sample group.

---

<sup>14</sup> CDM Report, page 6-45

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

		Variable IDs					
sampleky	sample group	4	8	39	42	58	59
105025	BS-08:8/23/2005:SW:S:-:-						
105178	BS-08:8/23/2005:SW:S:-:-	98	6.53				
105179	BS-08:8/23/2005:SW:S:-:-				0.134	0.5	0.014
106374	BS-08:8/23/2005:SW:S:-:-			68			
106848	BS-08:8/23/2005:SW:S:-:-						
105189	BS-117:9/14/2005:SW:S:-:-	132	10.18				
105190	BS-117:9/14/2005:SW:S:-:-				0.42	3.23	0.038
106079	BS-117:9/14/2005:SW:S:-:-						
106175	BS-117:9/14/2005:SW:S:-:-			13000			
106849	BS-117:9/14/2005:SW:S:-:-						

83. The rows of data for any given sample group are collapsed into only a single row. The rows were collapsed by moving values into missing areas in the first row of a given sample group. A sample group that has more than one value for a variable is averaged<sup>15</sup>. Below are the collapsed rows for the example given above.

		Variable IDs					
sampleky	sample group	4	8	39	42	58	59
106848	BS-08:8/23/2005:SW:S:-:-	98	6.53	68	0.134	0.5	0.014
106849	BS-117:9/14/2005:SW:S:-:-	132	10.18	13000	0.42	3.23	0.038

84. The sample keys (the first column) are not indicated in the SW3 Excel sheet because all of the sample keys have been collapsed into individual samples. We were able to match all of the EDA\_Samples in Appendix F with the collapsed sample groups **but not all of the values**. Dr. Olsen's SW3 Excel sheet has 573 rows with 26 variables; therefore his sheet has 14,898 values. His SW3 data has 915 missing values. The SW3 Excel sheet we produced has 573 rows and the same 26 variables, but the composition of the entries is very different.

<sup>15</sup> CDM Report, page 6-47

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

Dr. Olsen's SW3 data	13,983 values + 915 missing values = 14,898 values
<hr/>	
Analytic Focus	12,933 matched data values (Agreement with Olsen)
+	849 matched missing values (of the 915) (Agreement with Olsen)
+	499 missing values (Database is missing, but Olsen has data)
+	66 data values exist (Database has data, but Olsen has missing)
+	551 non-matched values (Database and Olsen's Excel Files differ)
=	14, 898 values

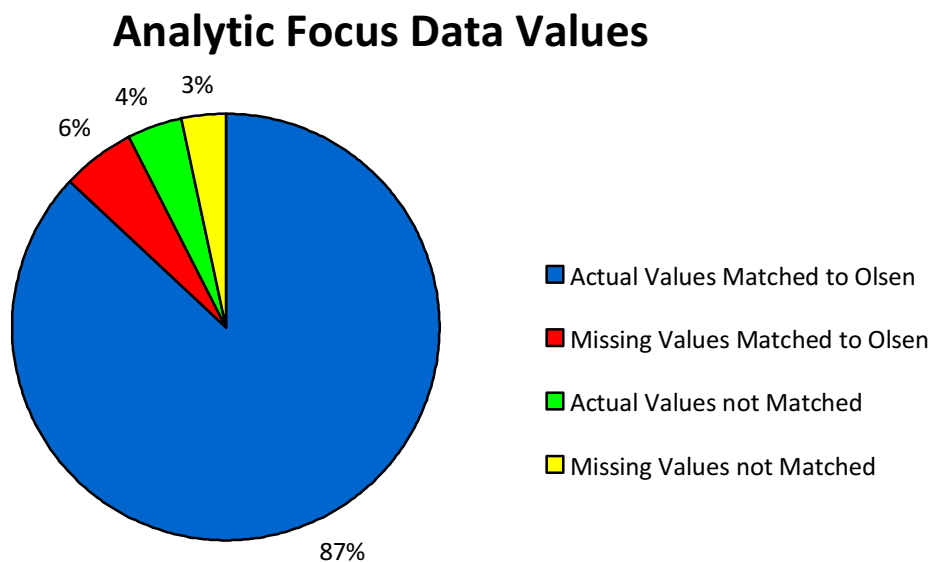
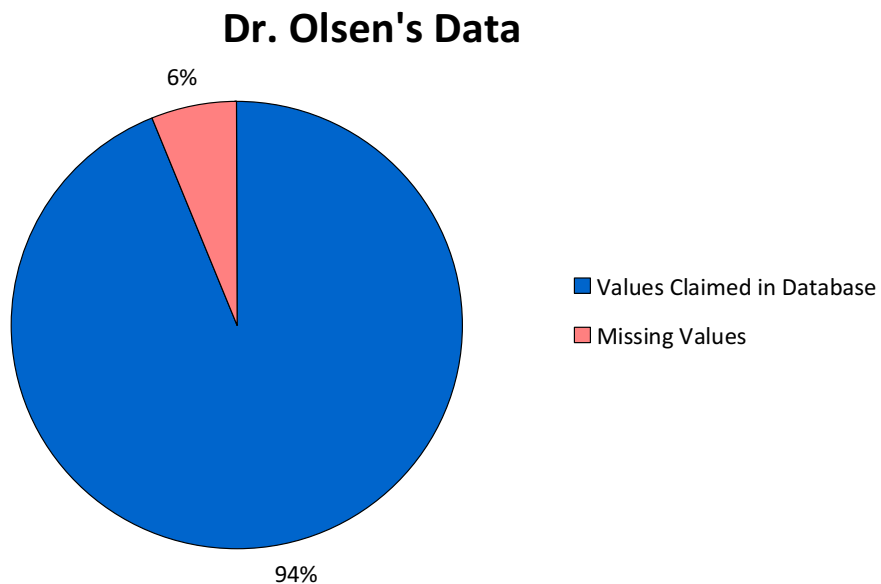
---

85. To summarize, of the 915 missing values that Dr. Olsen had, we found only 849 missing values – the remaining 66 were decreed by Dr. Olsen to be missing when they in fact had data. In addition, there are 499 additional values that were missing data in the Access database, but which suddenly have data in Dr. Olsen's analysis file. Finally, there are 551 values in the dataset where the value in the Excel file used for analysis differed from the original values in the Access database. In total, there are over 1,000 cells in Dr. Olsen's analysis database that do not correspond to the original data. This is about 7.5% of the total data that is in error or changed in some manner. This calls into question any quality of any analysis or data used by Dr. Olsen. Additionally, the 1,116 cells that have discrepancies are only one part of the problem. There is also a significant amount of data thrown away or ignored for no discernable reason.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

86. These outcomes are summarized in the next two charts.



REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

87. Dr. Olsen calculates his PCA scores in Appendix F of the CDM report in an Excel sheet (“To calculate a PC score for each individual sample, the PC coefficient is multiplied by the standardized parameter concentration. This is performed for all parameters (variables) (*sic*<sup>16</sup>) in a particular PCA run. The product values for all 25 (*sic*<sup>17</sup>) parameters are summed to yield one PC score for each sample for each PC. Hence, a particular sample will have both a PC1 and a PC2 score”).<sup>18</sup> We reproduced Dr. Olsen’s PCA scores in the following manner.

88. Start with the original SW3 data for the 26 variables. Missing values are replaced with the means of the variables before taking the logarithms. Compute z-transformations (subtract the mean of a variable, divide by it’s standard deviation) on these original variables. Multiply the SW3 z-transformed variables by the first two sets of coefficients produced from Dr. Olsen’s PCA on the SW3 log base ten data, ignoring the remaining sets of PCA coefficients. This produces two variables with 573 observations each. The 573 observations are the EDA\_Samples (S1, ....., S573). The two variables are PC1 and PC2.

89. To calculate PC1 for the first EDA\_Sample “S1”, find the minimum value of the PC1 column, take the absolute value of the minimum value, add 1 to this value, then add the value of the first EDA\_Sample. This method does not correspond to any standard PCA methodology.

---

<sup>16</sup> Dr. Olsen throughout his report confuses the terms parameter and variable. In this sentence he uses one to explain the other. From context, it seems that Dr. Olsen means variable when he says parameter. A parameter is a single value that describes a characteristic of a population, like an arithmetic mean or a variance. A variable is a theoretical construct used to denote a value that can change according to the sample being observed. These are not interchangeable terms.

<sup>17</sup> There are 26 variables in Dr. Olsen’s analysis, not 25.

<sup>18</sup> Dr. Olsen’s calculations are described on page 6-53 in the CDM report

---



REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

90. The calculations above in the previous two paragraphs used to duplicate the PCA values do not match the description of how to calculate PCA values given in the CDM report. The CDM report does not describe taking the absolute minimum of a column as a part of the calculation. Using this procedure, we were able to exactly replicate the scores used by Dr. Olsen.

91. In this process, Dr. Olsen commits an error so basic and so egregious that it completely invalidates every result and conclusion he offers. He runs the PCA on the logarithmic scores, but he ignores the SYSTAT program's calculations and instead applies the PCA coefficients to the original data without taking the logarithms.

92. Just to be clear, I will repeat the steps taken by Dr. Olsen for analysis:

- a. Take original data in the dataset with no missing data (26 variables)
- b. Take the logarithm base 10 of the values in the original data (26 new variables)
- c. Compute a transformation on the log values as (26 new variables again):

$$\text{New data value} = \frac{\text{Logged Data Value} - \text{Mean of Logged Data Values}}{\text{Standard Deviation of Logged Data Values}}$$

- d. Use the variables created in step c. to run the PCA in Systat
- e. Save the coefficients from Systat to apply to a different set of input data to compute scores for PC1 and PC2
- f. Compute scores for PC1 and PC2 outside of Systat using coefficients from step e. applied to data that skips step b. above.
- g. Translate scores for PC1 and PC2 from scale in step f. so that it appears there are no negative scores.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

93. **Systat readily computes the scores that Dr. Olsen wants.** However, Dr. Olsen ignores this and uses the coefficients from the analysis of the logarithmic data, but applies these coefficients to the original data without logarithms. What should have happened is that the coefficients should have been applied to the logged data to compute the scores.

94. To give a sense of the order of magnitude of this error, consider the problem of sending a rocket to Mars. The distance of Earth to Mars is a maximum of 250 million miles, which occurs when the planets are on the opposite sides of the Sun. On the log base 10 scale, the one used by Dr. Olsen, 250 million miles translates to 8.398. Remember that the logarithm computes the power of 10 needed to find the number of interest. So  $10^{8.398} = 250,000,000$ . Now compute fuel requirements on a distance of 8.4 miles (to be generous) and send the rocket off to Mars. The rocket would peak at just above 8 miles (not the 250 million needed) and then fall to Earth since it wouldn't even clear the atmosphere. This is the calculation that Dr. Olsen has done.

95. Dr. Olsen computes all of his coefficients on the logged data and then applies the coefficients to the original data ignoring the key transformation he has made. This should have been glaringly obvious when Dr. Olsen plotted his output for PC1 and PC2. The components computed are uncorrelated with one another – this is the entire basis for the computation of principal components, namely that each one is forced to be uncorrelated with all other components. The rotation methods Dr. Olsen uses enforce this – they force the rotated results to be uncorrelated, so whether one looks at rotated solutions or the unrotated solutions, they must be uncorrelated. The correlation between Olsen's PC1 and PC2 is  $R = 0.31$  when it should be identically zero.

96. The correlation between Systat's PC1 and PC2 is  $R = 0.0000000$ , just as it should be.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

97. This should be immediately obvious to any observer who knows anything about PCA.

Charts 9 and 10 on the next page present Dr. Olsen's score plots using his incorrect calculation and the correct score plots using the information from Systat, Dr. Olsen's program of choice.

98. Every conclusion that Dr. Olsen draws about his results that involve the use of the scores is wrong and meaningless. Computing the values that he did where there is confusion between the scales used means that Dr. Olsen had no idea what he was looking at and drew completely erroneous conclusions based on a mistake that he or his subordinates made. A quick check of the correlations between the scores would have immediately shown this error for what it was.

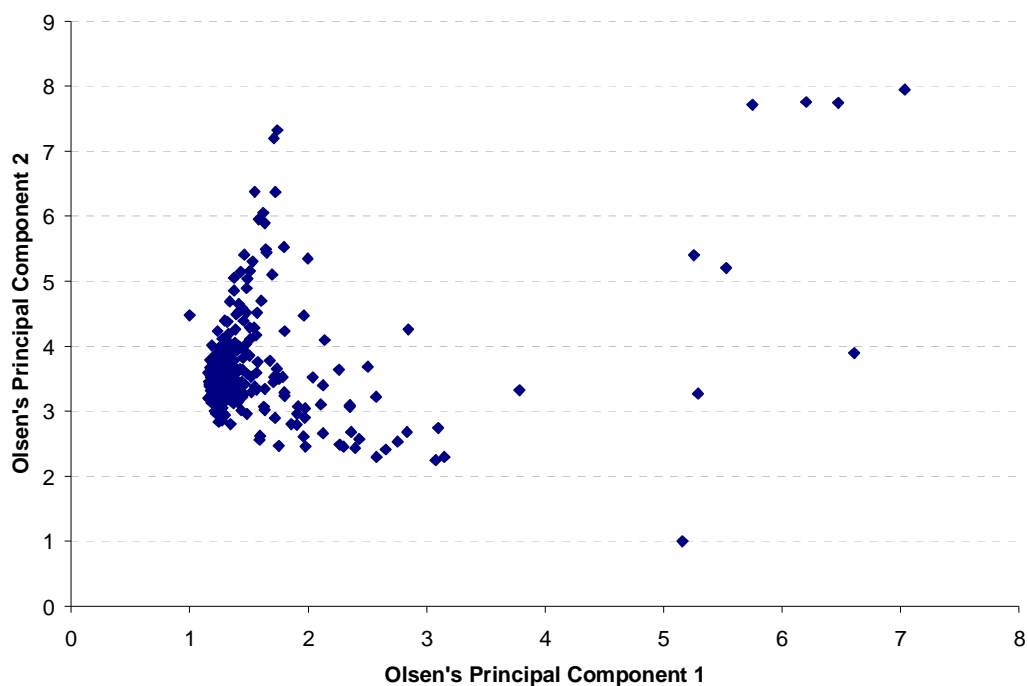
99. Finally, since the SYSTAT values are still on the logarithmic scale, the proper interpretation of the values would be on a real-world scale. This is easily done by computing the inverse logarithm of the Systat scores (raising 10 to the power of the score). This result is shown in Chart 11 below. When one examines this chart, one sees that there are a few extreme values charted on this plot – these result because Dr. Olsen didn't do quality control on the outliers in his data and so extremes result that are meaningless. On the proper scale, results appear on either PC1 or PC2, and there are four samples that result in extreme values in the center of the chart that are most likely due to quality control lapses.

100. Dr. Olsen's analysis, his charts, and his conclusions should be dismissed as erroneous and misleading.

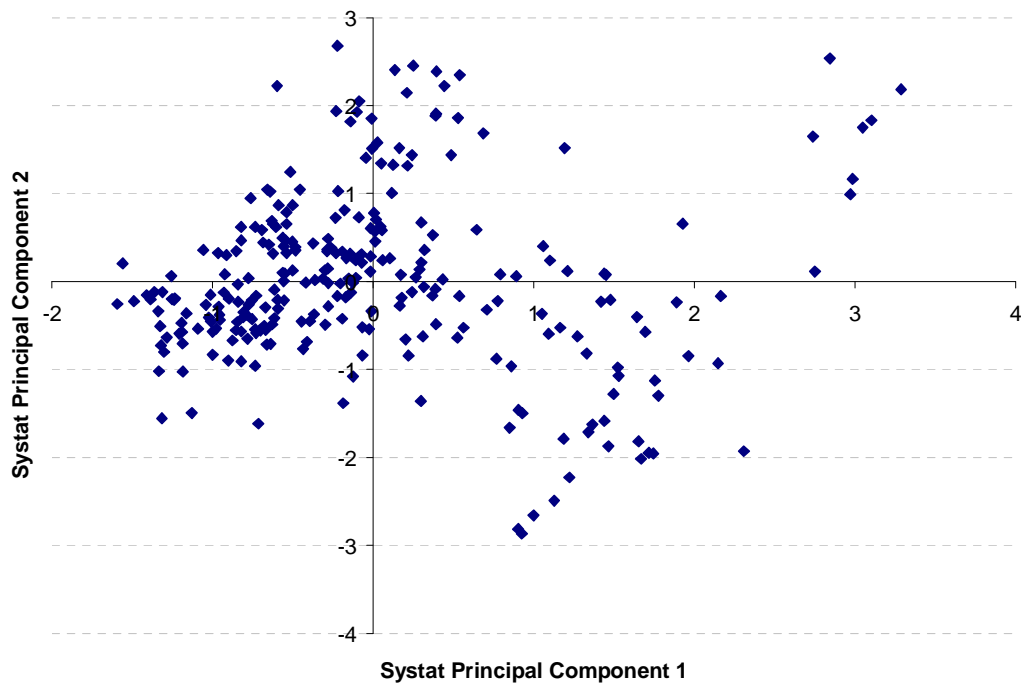
REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

**Chart 9: Olsen's PCA Score Plot**



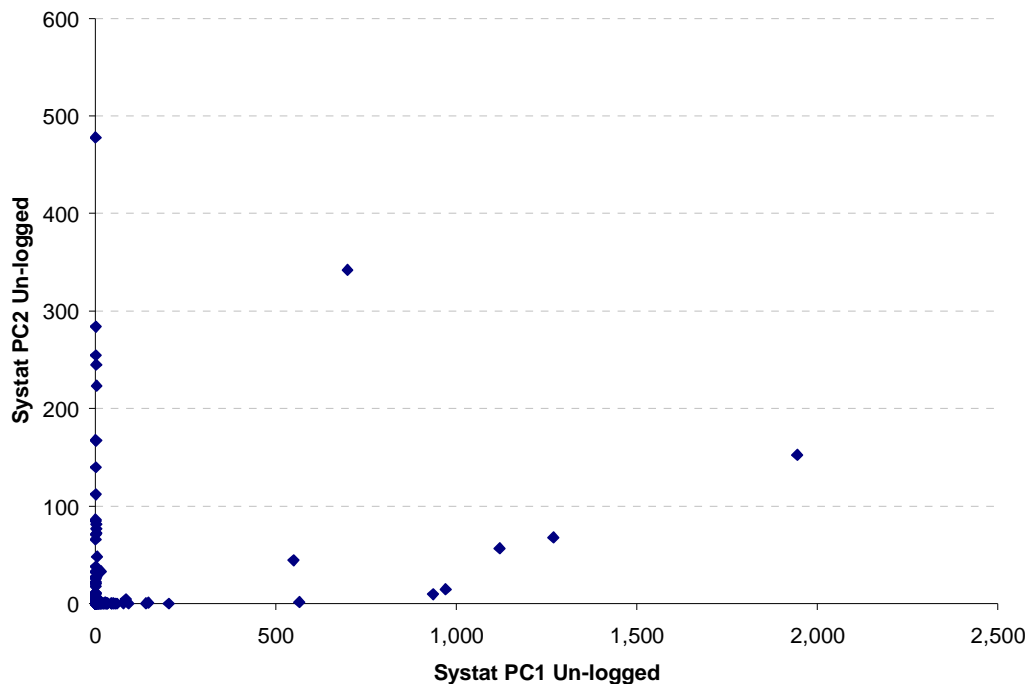
**Chart 10: Systat PCA Score Plot for Principal Components 1 and 2**



REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

**Chart 11: Systat Principal Components Converted from Logarithmic Scale to Real World Scale**



*Revisiting Missing Data: A File With 419 Samples And 56 Variables Without Missing Values*

101. We were able to use the original 315 variables found in Dr. Olsen's original database and create an Excel sheet with 419 samples and 56 variables with no missing values.

Remember that Dr. Olsen had only 267 samples with 26 variables. In our recreation of the Excel sheet, the variables were selected based only on percentage of observations available.

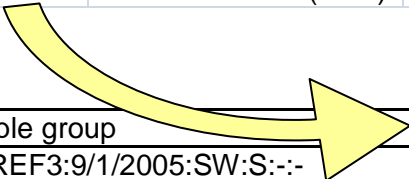
102. A query run on the database that downloads all of the records with "SW:S", which is all of the surface water data produces 66,260 rows of data. These are ported into an Excel sheet. A small example from this Excel sheet is presented below. This Excel sheet is then transformed into an intermediate Excel sheet that looks like the second example table below. This

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

intermediate Excel sheet has 6,564 rows of sample data with sample group and associated variable values.

103.

Sampleky	Paramky	ParamID	Value	SampleGrp
105152	4	Alkalinity (as CaCO3)	182	BS-REF3:9/1/2005:SW:S:-:-
105152	8	Chloride	12.44	BS-REF3:9/1/2005:SW:S:-:-
105153	42	Nitrite + Nitrate (as N)	0.368	BS-REF3:9/1/2005:SW:S:-:-



		Parameter Key			
sampleky	sample group	4	8	39	42
105152	BS-REF3:9/1/2005:SW:S:-:-	182	12.44		
105153	BS-REF3:9/1/2005:SW:S:-:-				0.368
106054	BS-REF3:9/1/2005:SW:S:-:-				
106172	BS-REF3:9/1/2005:SW:S:-:-			22	
106395	BS-REF3:9/1/2005:SW:S:-:-			42	
106871	BS-REF3:9/1/2005:SW:S:-:-				
102151	EOF07:5/15/2005:SW:S:-:-	402	30	340	0.277
106276	EOF07:5/15/2005:SW:S:-:-				
102232	EOF07:5/23/2005:SW:S:-:-				1.397
102233	EOF07:5/23/2005:SW:S:-:-				
104737	EOF07:5/23/2005:SW:S:-:-			3000	
106277	EOF07:5/23/2005:SW:S:-:-				

**Paramky 39** doesn't appear until much later in the data, thus it's absence from the first table.

104. Data from the database usually has several samples for each sample group (six samples for the BS-REF3:9/1/2005:SW:S:-:- sample group in above example). The rows of data for a given sample group are collapsed into a single row. The rows are collapsed by moving values into missing areas in the first row of a given sample group. When there is more than one value for a variable the values are averaged. Below are the collapsed rows for the above example.

sample	sample group	4	8	39	42
	BS-REF3:9/1/2005:SW:S:-:-	182	12.44	32	0.368
	EOF07:5/15/2005:SW:S:-:-	402	30	340	0.277
	EOF07:5/23/2005:SW:S:-:-			3000	1.397

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

105. This Excel sheet is the summary spread sheet for the creation of the database. This spreadsheet (referred to as EDA\_Sample in the CDM report) has 2,681 rows with 315 columns (each column is a variable).

106. This spreadsheet is further reduced by eliminating samples that have fewer than 20 observations on the 315 variables (i.e. of the 315 variables, only 19 or fewer have data). This results in an Excel sheet with 835 sample rows with at least 20 variable values in each row.

107. The database is further reduced by keeping only variables with < 24% of missing values. This produces an Excel sheet with 835 samples and 56 variables.

108. Finally, we retain only samples where all of the 56 variables have values (i.e. no missing values for the variables). This final Excel sheet has 419 sample rows with 56 variables and no missing values.

109.

<u>What I Did to Obtain the Maximum Available Data</u>	<u>Samples</u>	<u>Variables</u>
Full Data After Collapsing to Final Structure for Samples	2,681	315
<u>Samples with Less than 20 Variables with Data</u>	<u>1,846</u>	315
Samples with a Minimum of 20 Variables with Data	835	315
<u>Variables with data in less than 25% of samples</u>	835	<u>259</u>
Reduction to Variables with 25%+ Samples with Data	835	56
<u>Samples with Any Missing Data on 56 Variables</u>	<u>416</u>	56
Samples with No Missing Data on 56 Variables	419	56

<u>What Dr. Olsen Retained</u>		
Samples Dr. Olsen Retained	573	26
<u>Samples with Any Missing Data on 26 Variables</u>	<u>306</u>	26
Samples with No Missing Data on 26 Variables	267	26

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

110. In our reduction to the smallest dataset with no missing data, we obtained 56 variables. These 56 variables do NOT include the four bacteria variables. The Excel sheet created following Dr. Olsen's own methods is **missing four of the 26 variables** crucial to his PCA runs. The four missing variables are total coliforms, E. coli, enterococcus, and total coliforms.

111. The four bacteria variables were forced back into the analysis data set. This produced a large data set with 835 samples and 60 variables. However, when we then eliminate samples with missing data, we keep only **296 samples and 60 variables**. This is our final dataset for analysis, constructed considering only the use of all data and the elimination of samples with missing data.

112. There is no consistent explanation of the difference between Dr. Olsen's data and the data set we constructed. It is NOT accounted for with data rejected by Dr. Olsen because of his claims for samples in areas with cattle. There are, furthermore, severe differences between data on the Access database and data in Dr. Olsen's final Excel database, indicating that he: added data values in some cases with no documentation as to why, threw away data values in other cases, again with no documentation as to why, and on 3% of the records changed values with no explanation as to why.

113. Using the full set of data we derived with no missing values and including the bacterial data, we reanalyzed the data. Results are shown in Table 3 below.



## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

**Table 3. Analysis of the 296 Samples With 60 Variables.**

Rotated	1	2	3	4	5	6	7	8
TOTAL_CADMIUM	0.973	0.065	-0.031	0.006	0.016	0.064	0.046	0.089
TOTAL_BERYLLIUM	0.962	0.131	-0.012	-0.005	0.023	0.044	0.091	0.032
TOTAL_SILVER	0.943	0.045	-0.014	0.04	0.038	0.06	-0.007	0.166
TOTAL_ANTIMONY	0.909	0.131	0.058	0.122	0.057	0.038	-0.067	-0.034
TOTAL_THALLIUM	0.894	0.095	0.059	0.099	0.073	0.032	-0.002	-0.001
DISSOLVED_CADMIUM	0.865	0.001	0.469	-0.077	-0.001	0.044	-0.043	-0.009
DISSOLVED_THALLIUM	0.858	0.007	0.49	-0.068	-0.008	0.042	-0.036	-0.012
DISSOLVED_BERYLLIUM	0.851	0.009	0.466	-0.061	0.012	-0.004	-0.032	-0.009
DISSOLVED_ALUMINUM	0.828	0.209	0.355	-0.208	0.003	0.093	0.07	0.042
TOTAL_SELENIUM	0.81	0.16	0.08	0.124	0.063	0.075	0.068	-0.22
TOTAL_KJELDAHL_NITROGEN	0.791	0.412	0.021	0.024	0.069	0.093	0.046	0.006
DISSOLVED_ANTIMONY	0.744	0.06	0.615	-0.062	0.034	0.066	-0.049	0.056
DISSOLVED_IRON	0.732	0.299	0.481	-0.196	0.067	0.1	0.038	0.005
DISSOLVED_LEAD	0.726	0.122	0.594	-0.069	0.004	0.041	0.001	-0.064
DISSOLVED_SILVER	0.725	0.033	0.666	-0.069	0.036	0.028	-0.047	0.033
DISSOLVED_VANADIUM	0.717	0.029	0.515	-0.014	-0.004	0.065	-0.027	0.244
DISSOLVED_COBALT	0.604	0.286	0.515	0.059	0.048	0.053	-0.01	0.076
TOTAL_P_4500PF_	0.129	0.883	0.171	0.076	-0.203	0.181	-0.029	0.042
TOTAL_COPPER	0.146	0.877	0.07	0.037	0.083	0.162	0.003	-0.063
TOC	0.138	0.851	0.149	0.046	0.143	0.242	0.002	-0.034
TOTAL DISSOLVED_P_4500PF_	0.067	0.815	0.183	0.142	-0.335	0.191	-0.139	0.033
TOTAL_POTASSIUM	0.121	0.814	0.207	0.367	-0.152	0.071	-0.059	0.022
TOTAL_NICKEL	0.268	0.783	0.11	0.176	0.082	0.045	0.235	0.266
SOLUBLE_REACTIVE_P_4500PF	0.034	0.783	0.166	0.129	-0.36	0.222	-0.093	0.059
TOTAL_ARSENIC	0.342	0.762	0.102	0.135	0.17	0.121	0.134	0.107
AMMONIA_NITROGEN	-0.257	0.759	0.12	0.061	0.227	0.118	0.094	-0.091
TOTAL_IRON	0.238	0.692	0.035	-0.338	0.255	0.146	0.398	0.02
TOTAL_ALUMINUM	0.234	0.665	0.026	-0.386	0.151	0.175	0.423	0.052
TOTAL_ZINC	0.513	0.622	0.052	-0.002	0.093	0.102	0.153	0.1
TOTAL_COBALT	0.572	0.611	0.015	-0.079	0.067	0.046	0.341	0.158
DISSOLVED_BARIUM	0.133	-0.164	0.913	0.091	-0.147	0.054	0.15	-0.005
DISSOLVED_MAGNESIUM	0.232	0.185	0.898	0.17	0.036	0.024	-0.046	-0.098
DISSOLVED_CALCIIUM	0.138	-0.292	0.861	0.287	-0.001	0.026	0.037	-0.038
DISSOLVED_CHROMIUM	-0.133	0.131	0.814	0.009	0.11	0	-0.055	-0.016
DISSOLVED_SODIUM	0.051	0.136	0.796	0.433	-0.165	-0.132	-0.166	0.118
DISSOLVED_SELENIUM	0.564	0.049	0.764	-0.079	0.03	0.039	-0.035	-0.155

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

DISSOLVED_POTASSIUM	0.118	0.235	0.746	0.045	-0.221	-0.043	0.006	0.031
DISSOLVED_NICKEL	0.264	0.443	0.741	0.153	0.051	0.018	-0.065	0.193
DISSOLVED_ARSENIC	0.411	0.421	0.7	0.136	0.132	0.127	-0.059	0.045
DISSOLVED_COPPER	0.313	0.557	0.65	-0.084	0.025	0.138	-0.079	-0.091
DISSOLVED_ZINC	0.487	0.323	0.641	-0.017	-0.004	0.055	-0.097	0.02
DISSOLVED_MOLYBDENUM	0.6	0.178	0.606	0.036	0.029	0.015	-0.099	0.195
ALKALINITY_AS_CACO3	0.017	-0.208	-0.027	0.839	0.197	0.117	0.051	-0.144
TOTAL_CALCIIUM	-0.043	-0.396	-0.119	0.805	0.029	0.04	0.194	-0.046
TOTAL_SODIUM	-0.096	0.214	0.207	0.798	-0.215	-0.173	-0.179	0.19
CHLORIDE	-0.04	0.186	0.215	0.791	-0.235	-0.108	-0.141	0.188
TOTAL DISSOLVED SOLIDS	0.107	0.378	0.146	0.763	-0.071	0.006	0.117	0.056
TOTAL SULFATE_SO4	-0.106	0.343	0.215	0.676	-0.063	-0.162	-0.139	0.079
TOTAL_MANGANESE	0.104	0.581	0.118	-0.091	0.476	0.103	0.387	0.113
FECAL_COLIFORM	0.174	0.521	0.061	-0.058	0.039	0.772	0.053	0.042
E_COLI	0.102	0.533	0.045	-0.093	0.05	0.767	0.039	-0.022
TOTAL_COLIFORM	0.102	0.52	0.117	-0.023	0.016	0.713	0.032	-0.002
ENTEROCOCCUS_GROUP	0.192	0.502	-0.041	-0.094	0.057	0.688	0.109	0.122
TOTAL_BARIUM	-0.046	0.192	-0.19	0.068	-0.198	0.08	0.784	0.153
TOTAL_VANADIUM	0.112	0.103	0.038	0.121	0.143	0.065	0.134	0.822
TOTAL_LEAD	0.548	0.499	-0.009	-0.218	0.063	0.115	0.454	0.077
DISSOLVED_MANGANESE	0.183	0.305	0.599	0.024	0.41	0.051	0.157	0.056
NITRITE_NITRATE_AS_N	-0.185	-0.055	0.052	0.14	-0.774	-0.074	0.172	-0.136
TOTAL_MAGNESIUM	0.222	0.559	0.154	0.581	0.096	0.032	0.11	-0.142
TOTAL_CHROMIUM	-0.335	0.348	0.102	-0.012	0.141	-0.018	0.329	-0.144

## "Variance" Explained by Rotated Components

1	2	3	4	5	6	7	8
14.67	11.987	10.655	5.131	1.92	2.72	2.025	1.339
5				7	3		

## Percent of Total Variance Explained

1	2	3	4	5	6	7	8
24.45	19.978	17.758	8.552	3.21	4.53	3.376	2.231
8				2	8		

114. Many dissolved chemicals enter into the first and third principal components when they are included in the PCA runs. Phosphorus no longer loads onto the same component as fecal coliform, e-coli, total coliform and enterococcus. These latter do enter as a group to define a component, but not until the sixth principal component and not in the same component as the

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

phosphorus variables. These PCA results indicate that other variables are important in addition to the 26 variables discussed in the CDM report, and that the “signature” discovered by Dr. Olsen disappears when he brings in the full set of data available.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

**EXPERT REPORT OF VALERIE HARWOOD**

115. Dr. Harwood claims to have discovered a bacterium that carries a marker that uniquely identifies poultry litter – the poultry litter biomarker (PLB)<sup>19</sup>. Using this marker, she claims it is possible “to detect and quantify the amount of poultry-specific contamination in environmental samples, including soil, edge of field, surface water, and ground water samples collected in the IRW.”<sup>20</sup>

116. In her report, Dr. Harwood explains the importance of determining the sensitivity and selectivity of a biomarker. “Sensitivity (the frequency of positive results when the contaminating source is present) and specificity (the frequency of negative results when the contaminating source is absent) are among the most important attributes of a useful MST test.”<sup>21</sup> Failure to establish either of these characteristic renders a test useless, since high error rates mean that use of the biomarker is not reliable.

117. Dr. Harwood does not establish that the marker she claims to identify poultry litter has the requisite sensitivity or selectivity. Her sample is too small to measure any of the standards she establishes for either sensitivity or selectivity. Further, the cavalier sampling procedure used by the state to collect materials for testing is so flawed as to render it useless. This report addresses the statistical problems with Dr. Harwood’s work and shows that none of her conclusions can be supported.

---

<sup>19</sup> CDM Report, page 6-31

<sup>20</sup> Harwood Report, paragraph 43, page 17

<sup>21</sup> Harwood Report, paragraph 42, page 17

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*General Principals of Sampling*

118. There are four common sense issues to consider in drawing a sample. The first is determining what is to be measured. The second is the precision required for the estimate. The third is whether the sample is “representative”. The term representative simply means that a sample selected from a population can be used to draw inferences about that population. The fourth is how to sample the population.

**What to Measure**

119. Determining what is to be measured is harder than it may seem. In Dr. Harwood's case, she wanted to know two statistics – how frequently does the marker show up in poultry, and how frequently is it absent in other animals. The latter is much more difficult to measure, since you are looking for something that may or may not be there.

120. However, this is a very common problem in statistics. For example, in biostatistics and epidemiology health researchers frequently face the problem of detection. The problem of detection is that, when an event occurs, we need to know about the event. The Centers for Disease Control (CDC) monitors hospitals for outbreaks of unusual diseases or patterns of diseases. The EPA and NASA are working jointly to combine satellite data with EPA tower data to detect particulate air pollution. In quality control, the object of the collection of data is to find that flaws do not exist in the product or process. A similar process is used in accounting and audit. In all of these cases, a reasonable sample size is necessary to ascertain that the error rate is below a certain threshold.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

121. A simple example follows. Suppose I want to know how many people in the United States have a birthday in January. Assume that all months have equal numbers of days and that birthdays are spread uniformly throughout the year, so the chance of being born in January is one-twelfth. If I took a random sample of size 12, I'd expect to get one January birthday on average, but in fact I only have a 38% chance of getting one person with a January birthday. I have a 62% chance of getting zero or two. So if I take a sample of size 12 (a sample that would be large relative to what Dr. Harwood uses), I could extrapolate that result to the U.S. population, but I'd know very little about the number of people who have a birthday in January. In fact, I could easily conclude that NO ONE in the population has a birthday in January – but I'd be wrong. Now suppose that I want to know the number of people in the U.S. who have a birthday on January 1 in a non-leap year. The chance with the same small sample size of getting someone with a **rarer** event is much lower – in this case there is a 97% chance of getting no one with the January 1 birthday. But this does not mean that no one in the U.S. has a birthday on January 1.

122. But that's not the problem that Dr. Harwood is attempting to address. In the example above, I start with the conditions where **I know the probability** (because there are 12 months and a uniform distribution of births in my example). In the above example, I only want to know what the possible outcomes are. BUT Dr. Harwood doesn't know the relative frequency of the biomarker in poultry or more importantly in other species. For the example I gave above, where we know the probability is 1 in 12 (.083), we can determine how likely it is to get **no** positive outcomes. In the case where we don't know what the frequency of the biomarker is, we have to estimate the frequency from what we don't see. If we sample 12 animals using independent sampling techniques and observe no animals carrying the biomarker, the estimate for the proportion of animals with the biomarker is as high as 32% with a one percent chance the

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

number is actually higher than that. It could be 32% of animals in the total population carry the biomarker. It could be 21%, it could be 5% - we have no way of knowing.

123. Note that, Dr. Harwood doesn't have samples as large as 12 – hers are on the order of two or three. That means that the probability that animals in the population carry the biomarker can be as high as 78%, even when she doesn't observe an animal with the biomarker.

### **How Precise?**

124. In the examples given above, the discussion focused on one measurement – how many animals with a marker. Dr. Harwood compounds the difficulty of the problem by considering whether the rates she studies differ by whether they are in the IRW or outside.

125. For comparisons between groups, there is sampling variation associated with each group separately. There is a certain amount of sampling variation associated with samples from inside the IRW. There is also sampling variation associated with samples from outside the IRW. To compare the results between the groups, the amount of sampling variation would be expected to at least double if the sample sizes are the same, or more if the sample sizes are different.

126. From Dr. Harwood's Table 2<sup>22</sup>, she shows samples of size 5 for beef cattle inside and 5 for beef cattle outside the IRW. With a sample of size 5, the best she could determine is that with no samples showing the biomarker, the prevalence of the biomarker is less than 60%. In other words, as many as 60% of the cattle in either group could have the biomarker and she

---

<sup>22</sup> Harwood Report, page 24

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

could still get a sample of five cattle with no biomarker. As for comparing the rates between two groups of cattle, each with five in the sample, it would be impossible to make any determination of the differences between two groups.

127. In the same table, she shows that she has only one swine sampled inside and one sampled outside. There is **NO** statistical test that would allow any comparison of a sample of one with a sample of one; it is impossible to calculate any variability around the sample estimate. For a sample of size one, the outcome is either 0% (no marker) or 100% (marker found) – not very informative. The reliability of an estimate for a characteristic in the population is based on the variability measured in the sample. With a sample of size one, the variability is infinite, meaning there is no reliability. In other words, by taking a sample of one swine, Dr. Harwood knows nothing about whether any other swine in the population of swine carry the marker.

128. In fact, the same can be said for all sample sizes in this table, all of which are of size 5 or less. They are all so small that no inferences can be drawn from any of these samples about the presence or absence of the biomarker. Further, it would be impossible in most cases to even compare inside the IRW to outside the IRW since the sample sizes are too small to permit the calculation – it isn't even possible to do the calculation since the calculation in some cases would involve dividing by zero, a mathematical impossibility.



REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

129. Finally, the other issue of concern is the precision Dr. Harwood says is required for her results to be acceptable. From her first deposition, Dr. Harwood says that “generally, in PCR, your error rate should be 5 percent or less”.<sup>23</sup>

130. To calculate an error rate, one has to set up a table of the form:

Test	Truth	
	Positive	Negative
Positive	a	b
Negative	c	d

where the entries a, b, c, and d are counts of outcomes.

131. To measure a number like five percent, either  $b+c$  has to be very low relative to  $a+d$  (to ensure that the number is much lower than five percent), or the sample size has to be large enough to make a determination of the actual proportion. Consider that for a sample of size 31 (the number in Harwood’s Table 2), the only proportions that can be estimated are  $0/31$ ,  $1/31$ ,  $2/31$ , and so on, or 0%, 3.2%, 6.4%, 9.6%, and so on. In other words, if there is more than one error, the test fails the standard set by Dr. Harwood with this sample, and that requires combining all samples from cattle, swine, ducks, geese, and humans and the assumption that the error rates are the same for all groups. If they differ between groups, the problem is compounded since there is another level of variability in the outcomes to control.

132. To summarize, Dr. Harwood cannot achieve her own standard:

- the sample size is too small to even measure a number like 5%,
- with multiple types of animals tested, the measurement is confounded,

---

<sup>23</sup> Harwood Deposition 1, page 266, lines 10 and 11

- she did not perform these tests even though she clearly stated they were necessary for the acceptance of the biomarker.

## **Representativeness**

133. To ensure representativeness of a sample, there needs to be either a sampling procedure in place that catalogs where the population is to be found and how it will be sampled, or a model explaining the movement of the population and how it will be captured. For example, in a clinical trials setting, the sampling requires an assumption that the population is coming to hospitals and will be captured at random, and the real randomness in a clinical trial comes from the administration of a treatment or control to each person coming to the hospital.

134. In a experimental design setting, especially this one in the IRW, the population doesn't come to the researcher, so the researcher has to go to the population. There need to be standards as to how the sample is selected, the likelihood of selection of a unit in the sample, and methods to allow extrapolation of the results of the sample to the population.

135. Dr. Harwood has none of this in place. She doesn't know where the population is (e.g. dairy cows), how many there are, how they were selected, nor how the populations interact with one another. This latter point relates to the fact that many of the animals cohabitate in areas, and so are exposed to similar influences. Testing one group may be correlated with testing of another group, but Dr. Harwood wouldn't know this is the case since there is no sampling frame or documentation that describes how the sample was selected from the population, or even what the population is.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

136. This point is made over and above the concerns expressed above about the inadequacy of the sample sizes. Even with large sample sizes, the sample can be nonrepresentative (and biased) and so not very useful for analysis. In this case, no effort of any type was made to ensure that the sample represented a population.

### **How the Samples Were Selected**

137. The final point has to do with the method for sampling. Much of the “sampling” done in this case was cluster sampling. The analyses conducted by Dr. Harwood and all of the statistics she presents all rely on an assumption of simple random sampling. This simply didn’t happen.

138. A cluster sample is where multiple observations are taken from the same location, typically to reduce the cost of collection. The problem is that, units that are in the same location are more likely to be similar than units from different locations. This means that the variability of the observations increases, because there is the variability due to individual variation, and additionally the variation associated with the clusters being sampled.

139. This is a common phenomenon in surveys – a necessary evil to reduce the cost of data collection. In a survey of a human population about income, people who live on the same block are more likely to have similar incomes than those on different blocks. It is common to sample four households per sampled block in a survey to measure income or unemployment, but once the interviewer has spoken to the first household, there is less useful information from the other three households on the same block relative to households on other blocks.

140. The same is true for Dr. Harwood’s data collection. Sampling multiple cow pies within the same field leads to two problems. One is that the cows milling about together in the same

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

field are likely to be much more similar to one another than cows in different fields (same feed, same water supply, etc.). The second problem is even worse – what’s to keep the researcher from collecting multiple cow pies from the same cow? It’s not as if they are distinguishable on examination – unless you are following each cow.

141. Because of this, Dr. Harwood's data is even less “precise” than she would expect, since the sampling methods used lead to significant increases in the variability of the outcomes. For this and for all the preceding reasons, Dr. Harwood's findings are meaningless.

REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

**APPENDIX 1: PAST EXPERIENCE**

My background covers 30 years of research and study in the areas of statistics, economics, and their application to business problems. I am Managing Partner of Analytic Focus LLC, a company headquartered in Birmingham, Alabama. A portion of our work is conducting research in legal issues, including providing litigation support and expert witness services when requested. Some of our work focuses on measurement and mitigation of risk for financial intermediaries. The final area of our practice is in support of Federal and State agencies needing economic and financial analysis to pursue their missions. I am also a research professor in the School of Business and the School of Public Health at the University of Alabama – Birmingham.

In litigation, our firm has focused on class certification issues, intellectual property, antitrust, and regulatory compliance. In banking and insurance, we offer services regarding audit reliability, risk measurement, model validation, and optimization of operations. For regulatory agencies, we have contracts with several Federal agencies to determine risk to funds managed by the agencies or operations conducted by the agencies.

Prior to founding Analytic Focus, I was a director for ARPC, a firm in Washington, DC where I provided many of the same services currently offered by Analytic Focus. From the beginning of 1997 through the end of 1999, I was a Director for Price Waterhouse and subsequently PricewaterhouseCoopers. In this position I headed up two different staff groups, one the financial research group in the Survey Research Center (SRC) run by Price Waterhouse, the other the data mining group. Our research efforts in the SRC was in support of business to business consumer research and financial analysis and for the Federal Government to research regulatory impact. In the data mining group we provided fraud detection services for financial services organizations, optimization research for businesses concerned with supply chain issues in production, and analysis of delivery systems for a number of major delivery companies. This latter work required coordination with our supply chain group and the three directors in charge of these operations formed an “Analytical Trust” where we worked jointly on the statistical and financial aspects of the design for these programs. On the whole, we combined resources in this small group in operations research, statistics, mathematical economics, finance, and system design to answer complex analytical questions.

Before joining Price Waterhouse, I was Chief Statistician for the Federal Deposit Insurance Corporation and the Resolution Trust Corporation, where I was responsible for all research on valuation of properties and assets taken in by the FDIC and RTC in the banking crisis of the 1980s and 1990s. I also supported research into fraud, optimization of contracts with servicing companies, and consumer perceptions of their interactions with banks and savings and loans. I prepared and jointly presented results on the FDIC’s consumer research to Congress, specifically the House Banking Committee in hearings on how consumers perceive what they are told regarding retail transactions in banks.

During this time, 1991 to 1996, I also served on a number of independent review committees for different Federal agencies to evaluate the quality of research conducted or research proposed for the National Institutes of Health, for the Department of Health and Human Services, for the Department of Justice, for Treasury, and for the Department of Agriculture. These committees were formed specifically to determine how to determine whether research presented to the Federal government could support conclusions drawn or to consider whether research proposed in grant applications would be adequate to study the topic in question.

## REBUTTAL REPORT

REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

I also worked for a time in the private sector as Chief Statistician and a Vice President for Opinion Research Corporation, from 1989 through 1991. In this position I helped design over 100 consumer research studies focusing on acceptance of new products, pricing, and customer satisfaction. In particular I helped to design the largest ongoing customer satisfaction study conducted in the United States for the U.S. Postal Service to investigate all aspects of consumer reactions to operations of and interactions with the Postal Service.

From 1986 through 1989, I was the first Chief Statistician for the newly founded National Center for Education Statistics, an agency within the Department of Education. As the Chief Statistician I was responsible for the design of all surveys and research conducted by NCES, reports to Congress on the state of education in the U.S. and in the world, and on staff development in research methods. In particular, under my guidance, NCES was one of the first Federal statistical agencies to publish standards for operations and research. These standards are still required for the conduct of research by all NCES staff and all contractors working with the NCES.

From 1975 through 1986 I held a variety of positions at the U.S. Bureau of the Census, including Chief of the Survey Design Branch, where I was responsible for the technical aspects of all research conducted on the evaluation of surveys and the 1980 Decennial Census. I also designed research studies on the validity of surveys conducted by the Census Bureau, experiments to measure response validity, and helped a number of countries develop research programs regarding their economic and demographic research programs.

My first positions after graduation were with the Institute for Social Research at the University of Michigan and as Manager of the Survey Research Center at Oregon State University.

During this time I served on a number of different committees in professional associations including the American Statistical Association, the American Association for Public Opinion Research, and the Research Industry Coalition, including the presidency of the latter. For each of these associations I was involved in issues of ethics and professional standards in the research community.

I have also served as an adjunct or visiting professor at a number of universities, besides my current positions as an adjunct at UAB. I have also been an adjunct professor teaching statistics at the George Washington University and a visiting research professor at the University of Illinois.

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

**APPENDIX 2: RESUME OF CHARLES D. COWAN***Key Qualifications*

Charles D. Cowan is Managing Partner of ANALYTIC FOCUS<sub>LLC</sub>. Dr. Cowan has 30 years of experience in statistical research and design. He consults for numerous public and private sector entities on the design, implementation, and evaluation of research and the synthesis of statistical and sampling techniques for measurement.

Dr. Cowan has designed some of the largest and most complex research programs conducted by the Federal Government, including the Post Enumeration Program conducted by the Bureau of the Census to evaluate the 1980 Decennial Census, the Economic Cash Recovery valuations conducted by the Resolution Trust Corporation in 1990-95, and many evaluation studies conducted for the Justice Department, the Department of Defense, the Department of Housing and Urban Development, and the Treasury Department. He has provided expert advice to corporations and government agencies on the incorporation of complex research designs in demographic and economic measurement problems, including:

- Development of procedures used by the Resolution Trust Corporation and the FDIC for determination of the value of all assets held by the RTC\FDIC taken from failed banks and S&Ls. Results from this research were used in quarterly reports to Congress on the loss to the American taxpayer that resulted from these failures. These estimates of anticipated recoveries on assets were also used by the RTC and FDIC for financial reporting, leading these agencies to their first clean opinions from the GAO in their annual review of agency financial statements.
- Establishment of audit and sampling methods to determine the completeness and reliability of reporting and record systems. These procedures were used to both expand and streamline bank examinations for safety and soundness and also compliance measurement for the FDIC. These sampling techniques are applied in the audit of Federal agencies concerned with regulatory review of operations and systems, and related systems for banks, regulatory agencies, and law firms;
- Application of econometric and biometric procedures for measurement of credit risk in large portfolios of loans. These models are frequently used for a variety of purposes within financial institutions, such as the pricing of loans, the management of customers long term, decision making on workouts for delinquent loans, and for establishment of economic and regulatory reserves.
- Evaluation of research conducted for the Department of Defense, for the National Institutes of Health, and for the Department of Agriculture, each in response to Congressional inquiries on the validity of published results, and also for defendants in lawsuits involving evidence proffered by plaintiffs in furtherance of their suit.
- Model fitting and development of projection methods to measure the likelihood of loss or errors in recording in loans held by banks or put up for auction; measurement of the likelihood of fraud and/or noncompliance in systems, including bank holding companies, trading activities for brokers, and systems for compliance with health department and judicial requirements;

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

- Incorporation of population demographic models with financial assessment models to predict risk for insurance companies and corporations in terms of number and value of potential claims in mass tort litigation.
- Development of procedures used by the Bureau of the Census for apportionment of population for revenue sharing purposes and the estimation of the undercount in the Decennial Census of Population and Housing. These procedures include application of capture-recapture methods to measure the size of the undercount in the decennial census, use of network sampling as an alternative measure for population size, and measurement of the reliability of data collected in the Census.
- Development of statistical methods to quantify the size of populations, including nomadic populations for the Census of Somalia, the undercount and overcount in the Census of Egypt, the number of missing children in Chicago, IL, and the number of homeless persons and families needing services in several large cities with transient populations.

Dr. Cowan teaches graduate and undergraduate courses in survey methods, statistics, and computer methods for analysis. He is the co-author of two books, one on evaluation of survey and census methods and one on econometric measures related to the welfare of the U.S. economy. He has written numerous articles on statistical methods, sampling, rare and elusive population research, and optimization techniques.

Prior to cofounding ANALYTIC FOCUS<sub>LLC</sub>, Dr. Cowan was a Director with ARPC and with Price Waterhouse, where he specialized in financial research, survey research, and audit sampling. From 1991 to 1996, Dr. Cowan was the Chief Statistician for the Resolution Trust Corporation and the Federal Deposit Insurance Corporation, where he designed research necessary to measure the loss from the Savings & Loan Crisis of the late 1980's and capitalization requirements for the RTC funds from the U.S. Treasury. Dr. Cowan also served as the Chief Statistician for the U.S. Department of Education, where he designed large-scale surveys of educational institutions to measure resource needs and availability, and for Opinion Research Corporation, where he designed predictive models of demand for automobile manufacturers, banks, and large horizontally diverse firms like GE and AT&T. Dr. Cowan worked for the U.S. Bureau of the Census, where he was the Chief of the Survey Design Branch and developed many of the techniques in use today for the evaluation of coverage in surveys and censuses.



REBUTTAL REPORT  
REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*Education*

Ph.D., Mathematical Statistics, The George Washington University, 1984  
M.A., Economics, The University of Michigan, 1973  
B.A., English and B.A., Economics, The University of Michigan, 1972

*Professional Experience*

Co-Founder, ANALYTIC FOCUS LLC, January, 2002 to present.  
Director, ARPC, November, 1999 to December, 2001.  
Director, PricewaterhouseCoopers LLP, January 1997 to November, 1999.  
Chief Statistician, Federal Deposit Insurance Corporation / RTC, 1991 to 1996.  
Chief Statistician, Opinion Research Corporation, 1989 to 1991.  
Chief Statistician, National Center for Education Statistics, US Dept. of Education, 1986 to 1989.  
Bureau of the Census: Assistant Division Chief, International Statistical Programs Center, 1984 to 1986; Staff Liaison for Statistical Litigation Support, 1983 to 1984; Chief, Survey Design Branch, Statistical Methods Division, 1978 to 1983; Acting Chief, Survey Analysis and Evaluation Branch, Demographic Surveys Division, 1976 to 1978; Office of the Chief, Statistical Research Division, 1975 to 1976  
Survey Research Center, Oregon State University: Manager, 1974 to 1975  
Institute for Social Research, U. of Michigan: Assistant Study Director, 1972 to 1974.

*Professional Associations*

Adjunct Full Professor, Statistics, University of Alabama – Birmingham, 2002-present.  
Associate Professor, Statistics, George Washington University, 1993 - 1998.  
Visiting Research Professor, Survey Research Laboratory, U. of Illinois, 1983 - 1989.  
Consultant, Dept. of Community Psychiatry, Johns Hopkins U., July 1985 - Dec 1987.

*Professional Societies – Memberships*

American Statistical Association (ASA)  
American Association for Public Opinion Research (AAPOR)  
International Association of Assessment Officers

*Professional Societies - Positions*

President, Research Industry Coalition, 1999-2000  
Council Member, Research Industry Coalition, Representative from ASA, 1995-2000  
President, Washington/Baltimore Chapter of AAPOR, 1998  
Program Chair, American Association for Public Opinion Research, 1991-2  
Program Chair, Section on Survey Research Methods, ASA, 1989-90  
Secretary-Treasurer, AAPOR, 1985-1986  
Associate Secretary-Treasurer, AAPOR, 1984-1985  
Editorial Board, Public Opinion Quarterly, 1980-1984  
Editorial Board, Marketing Research, 1989-2000  
Chair, Conference Committee, AAPOR, 1982-1989  
Chair, Committee on Privacy and Confidentiality, ASA, 1980-1981

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

*Publications*

- Strumpel, Burkhard; Cowan, Charles; Juster, F. Thomas; and Schmiedeskamp, Jay; editors, Surveys of Consumers 1972-73, Contributions to Behavioral Economics, Ann Arbor: The Institute for Social Research, 1975.
- Duncan, Greg, and Cowan, Charles D., "Labor Market Discrimination and Nonpecuniary Work Rewards" in Surveys of Consumers 1972-73, Contributions to Behavioral Economics, Ann Arbor: The Institute for Social Research, 1975.
- Curtin, Richard T. and Cowan, Charles D. "Public Attitudes Toward Fiscal Progress" in Surveys of Consumers 1972-73, Contributions to Behavioral Economics, Ann Arbor: The Institute for Social Research, 1975.
- Cowan, Charles D., and Spoeri, Randall K., "Statistical Distance Measures and Test Site Selection: Some Considerations", Proceedings of the Computer Science and Statistics: Eleventh Annual Symposium on the Interface, 1978.
- Bushery, John R., Cowan, Charles D., and Murphy, Linda R., "Experiments in Telephone-Personal Visit Surveys", Proceedings of the American Statistical Association, Section on Survey Research Methods, 1978.
- Spoeri, Randall K., and Cowan, Charles D., "On the Use of Distance Measures in Test Site Selection: A Practical Application Using Census Data", Proceedings of the American Statistical Association, Section on Business and Economic Statistics, 1978.
- Hogan, Howard, and Cowan, Charles D., "Imputations, Response Errors, and Matching in Dual System Estimation", Proceedings of the American Statistical Association, Section on Survey Research Methods, 1980.
- Schwartz, Sidney H., Cowan, Charles D., and Sausman, Kenneth R., "Optimization in the Design of a Large-Scale State Sample", Proceedings of the American Statistical Association, Section on Survey Research Methods, 1980.
- Cowan, Charles D., "Modifications to Capture-Recapture Estimation in the Presence of Errors in the Data" presented at the meetings of the American Statistical Association, Biometrics Section, 1982 (no proceedings).
- Cowan, Charles D. "Interviews and Interviewing", The Social Science Encyclopedia, Routledge and Kegan Paul, Publishers, The Netherlands, 1984.
- Wei, L. J. and Cowan, Charles D. "Selection Bias", Encyclopedia of Statistical Science, John Wiley and Sons, New York, N.Y., 1984.
- Cowan, Charles D. and Malec, Donald J. "Capture-Recapture Models When Both Sources Have Clustered Observations", Journal of the American Statistical Association, June 1986, Vol. 81, # 394, pp. 347-353, and Proceedings of the American Statistical Association, Section on Survey Research Methods, 1984.

## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Cowan, Charles D. The Effects of Misclassification on Estimates from Capture-Recapture Studies. Unpublished doctoral dissertation, The George Washington University, September 1984.

Cowan, Charles D. "Misclassification of Categorical Data", Proceedings of the American Statistical Association, Section on Survey Research Methods, 1985.

Cowan, Charles D., Biemer, Paul P., Magnani, Robert J., and Turner, Anthony G., Evaluating Censuses of Population and Housing, Statistical Training Document, ISP-TR-5, U.S. Department of Commerce, Bureau of the Census, 1985.

Cowan, Charles D., Turner, Anthony G., and Stanecki, Karen "Design of the Somali Post Enumeration Survey (1986-1987)", Proceedings of the American Statistical Association, Section on Survey Research Methods, 1986.

Cowan, Charles D., Breakey, William R., and Fischer, Pamela J. "The Methodology of Counting the Homeless", Proceedings of the American Statistical Association, Section on Survey Research Methods, 1986.

Cowan, Charles D. and Malec, Donald J. "Sample Allocation for a Multistage, Multilevel, Multivariate Survey", Proceedings of the Fourth Annual Research Conference (ARC IV), U.S. Bureau of the Census, 1988.

Frey, Carolin M., McMillen, Marilyn M., Cowan, Charles D., Horm, John W., and Kessler, Larry G.. "Representativeness of the Surveillance, Epidemiology, and End Results Program Data: Recent Trends in Mortality Rates", Journal of the National Cancer Institute, Vol. 84, No. 11, June 3, 1992.

Cowan, Charles D., Breakey, William R., and Fischer, Pamela J. "The Methodology of Counting the Homeless, A Review" in Homelessness, Health, and Human Needs. Institute of Medicine, National Academy Press, National Academy of Sciences, Washington, D.C., 1988.

Cowan, Charles D., "Standards for Statistical Surveys in the Federal Government: Practices in the Center for Education Statistics", Proceedings of the American Statistical Association, Section on Survey Methods Research, 1988.

Sudman, Seymour, Sirken, Monroe G., and Cowan, Charles D., "Sampling Rare and Elusive Populations", Science, Vol. 240, pp. 991-996, May 20, 1988.

Cowan, Charles D., "Mall Intercepts and Clinical Trials: The Philosophy of Inference from Different Types of Research Designs" in Marketing Research: A Magazine of Management & Applications, Vol. 1, No. 1, March 1989.

Cowan, Charles D., "Mall Intercepts: Principles of Design for Research" in Proceedings of the Seventh Annual Advertising Research Foundation Research Quality Workshop, September, 1989.

Cowan, Charles D., "Estimating Census and Survey Undercounts Through Multiple Service Contacts" in Housing Policy Debate: Counting the Homeless: The Methodologies, Policies, and Social Significance Behind the Numbers, Volume 2, Issue 3, pp. 869-882, 1991.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

Cowan, Charles D., "Ratio vs. Regression Estimators in a Large Scale Survey of S&L's" in Proceedings of the Section on Survey Research Methods, American Statistical Association, 1992.

Cowan, Charles D., "A Longitudinal Survey and Reality Check for the Value of Financial Assets" in Proceedings of Statistics Canada Symposium 92: Design and Analysis of Longitudinal Surveys, November 1992.

Cowan, Charles D., and Wittes, Janet, "Intercept Studies, Clinical Trials, and Cluster Experiments: To Whom Can We Extrapolate?" in Controlled Clinical Trials, Vol.15, pp.24-29, 1994.

Cowan, Charles D., and Klena, Matthew K. "Use of the EM Algorithm for Allocation of Proceeds from Auctions and Bulk Sales" in Proceedings of the Section on Business and Economic Statistics, American Statistical Association, 1995.

Cowan, Charles D., "Coverage, Sample Design, and Weighting in Three Federal Surveys" in Journal of Drug Issues, October, 2001.

Cowan, Charles D., "Use of Mass Appraisals in Toxic Tort Litigation Involving Loss of Value" in Proceedings of the International Association of Assessment Officers, October, 2002.

Cowan, Adrian M. and Cowan, Charles D., "Default Correlation: An Empirical Investigation of a Subprime Lender", The Journal of Banking and Finance, March 2004.

Cowan, Charles D. and Cowan, Adrian M., "A Survey Based Assessment of Financial Institution Use of Credit Scoring for Small Business Lending", SBA Report 283, Nov. 2006

Keith, Scott W. , Wang, Chenxi, Fontaine, Kevin R. , Cowan, Charles D. and Allison, David B. , "Body Mass Index and Headache Among Women: Results From 11 Epidemiologic Datasets", Obesity, Volume 16, Issue 2 (February 2008) 16: 377-383; doi:10.1038/oby.2007.32

Cowan, Charles D., and Cowan, Adrian M., "Quasi-Likelihood Estimation of Loan Portfolio Defaults in the Presence of Default Correlation and Autocorrelation", The European Journal of Finance (forthcoming, 2009)

Brock, David W., Thomas, Olivia, Cowan, Charles D., Hunter, Gary R., Gaesser, Glenn A., and Allison, David B., Association between Physical Inactivity and Prevalence of Obesity in the United States, Journal of Physical Activity and Health, (forthcoming, January, 2009)

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

### APPENDIX 3: PAST TESTIMONY

#### *Trademark Infringement:*

Quiksilver v. Brunswick, circa 1997. Deposed, case settled. Worked for the defendant, who had started producing t-shirts under brand name Quiksilver, one of their boat lines. The boat line could be named Quiksilver, but Quiksilver produces “surfer” clothes and were concerned about trademark confusion. We conducted a survey to determine level of confusion and the likely damages caused. Brunswick dropped the t-shirt line and settled.

St. Johns Knits versus St. Johns, circa 1997. Deposed, case settled. Small firm in California named itself St. Johns and began to produce ladies casual apparel with name of St. Johns. Worked for plaintiff, conducting survey on trademark confusion and calculating damages.

Nitro Leisure Products v. Acushnet. **Antitrust**, Trademark, and Deceptive Sales Practices filed in Florida. Deposition in 2003, settled in 2004. Worked for defendant. Issue was whether claims regarding the performance of “used and repackaged” golf balls were valid. Survey conducted, used to support damage claims. Second simultaneous suit was Acushnet v. Nitro – work used in settlement of the two simultaneously.

Community First Bank v. Community Banks. Trademark infringement. Deposition, October, 2004. Worked for Defendant. Issue was that Pennsylvania based Community Banks, a multi-state bank, opened branches in Northern Maryland. Community First Bank claimed it already had a charter in Maryland and the intrusion of Community Banks diminished the value of their name. Case resolved in favor of Defendant – dismissal on Summary Judgment.

#### *Trade Dress*

Sound Board Manufacturer v. European Manufacturer. Trade dress infringement. Worked for plaintiff. Circa 1997. Issue was that European manufacturer bought a sound board from U. S. manufacturer, reverse engineered it, and sold their copy with exactly same layout and design in competition with U.S. manufacturer. Conducted survey of bands, churches, small recording studios, and other potential purchasers of mid-price sound boards. Case settled.

Guntersville Breathables v. Kappler. Trade dress infringement. Worked for plaintiff. 2004. A manufacturer of camouflage hunting clothes developed a unique camouflage design and used it for their primary line of clothes. A second manufacturer bought materials from same fabric company and produced exactly the same design for hunting clothes sold in similar outlets to the same population of hunters. Survey designed and implemented. Case settled.

VPX v. ABB. Trade dress infringement. Worked for defendant. 2006-7. A manufacturer of liquid energy drinks sold in gyms, health clubs, and big box retailers filed suit against another manufacturer of liquid energy drinks, claiming that the shape and type on the bottles of the defendant were the same as that of the plaintiffs and caused confusion among potential purchasers. Conducted surveys of potential purchasers of liquid energy drinks to determine whether confusion exists. Deposition in January, 2007, testimony in bench trial in January, 2007. Defendant won.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*Patent Infringement:*

Smith & Nephew v. Zimmer, circa 1999. Deposed, then case settled. Worked for defendant who admitted infringing on patent but claimed that the particular feature upon which they infringed was not important to the choice of the product by physicians. Product was hip replacement “cup” and “stem”, and feature was machining of cup to minimize friction. We conducted a survey of physicians to determine what features were important to the selection of a hip replacement part. We used the survey to calculate damages; results were used in the settlement deliberations.

*Design Patent Infringement:*

Leatt v. Alpinestar. Worked for Plaintiff. Leatt has a design patent on a neck brace used in active sports like Motocross and has successfully defended their patent against imitators. Alpinestar produced a new neck brace allegedly based on the Leatt design. Conducted a survey to establish whether confusion existed regarding whether the designs between the two neck braces were considered to be the same or substantially the same. Case on-going.

*Deceptive Sales Practices:*

Executec v. Appleton Papers, circa 1998. Deposed, testified at class certification hearing. Class denied. Issue was whether Appleton Papers colluded with other manufacturers in the pricing of thermal fax paper products. Appleton had already won an antitrust case in Federal court on same issue. Conducted survey of pricing of product throughout Florida and proved that pricing of product was so discretionary at retail level that it was impossible to consider whether producer pricing had claimed impact at retail level. Case cited by Third District Court in Florida when tobacco class ruling in Florida was overturned on appeal.

Watkins et al. v. Dry Cleaners International, 2003. Not deposed, case settled before class hearing. Worked for defendant. Issue was whether DCI had properly informed customers of surcharge imposed to cover environmental costs. Plaintiffs claimed customers were confused and thought charge was improperly imposed tax. Survey conducted, damages calculated.

Fidelity Mortgage v. Seattle Times. Deceptive Trade Practices in Seattle Washington. Deposition in 2004. Worked for plaintiff. Damages calculated on lost sales because of publication of false interest rates. Case in appeals court.

Irena Medavoy v. Arnold Klein, M.D. et al.. Deceptive Sales practices case in California involving Botox, representing the cosmetics manufacturer. Worked for defendant. Deposed in 2004, case dismissed.

*Disparate Impact \ Discrimination*

Apkins et al. v. Atlantic Marine – Mobile. Loss of jobs, loss of work hours, lack of promotions for population of blacks working for a manufacturer who laid off blacks first, re-hired (called back) blacks last, refused to promote, and kept overtime for only certain workers. Worked for plaintiffs. Analysis of hiring practices, lay off records, filings with Federal government, and other records to develop pattern of practice analysis. Case settled, no testimony.



## REBUTTAL REPORT

## REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

Disparate impact in promotions for minority workers for a large public utility. Worked for plaintiffs. Analysis of testing and promotion procedures, development of methods to ascertain if skill tests used led to disparate treatment of minorities. Report submitted, case ongoing.

HOPE v. Illinois Chinese American Residence for the Elderly. Disparate impact for senior citizens for a public housing authority. Worked for city housing authority – plaintiff. Survey of senior citizens in a city to determine their attitudes and beliefs regarding different Federally sponsored senior citizen independent living facilities. Analysis of demography of general population in the city and comparison to distributions of residents in all independent living facilities in the city. Report and affidavit submitted, case ongoing.

Stein et al v. SLG Group. Disparate impact for minorities in availability of cemetery plots in multiple cemeteries owned by single holding company under the Fair Housing Act. Analysis of sales of plots to individuals to ascertain whether a pattern of practice existed. Worked for defendant. Case settled.

AHF COMMUNITY DEVELOPMENT v. City of Dallas. Disparate impact for minorities and families under the Fair Housing Act. Code inspections by police in the City of Dallas allegedly caused disruption and loss of fair use of housing in an affordable housing complex. Analysis of business reasons under HUD guidelines for all code inspections conducted by police and analysis of discriminatory and disparate impacts on residents. Deposition, July 2008. Case ongoing.

Mississippi Home Builders v. City of Brandon, Mississippi. Disparate impact for minorities and families under the Fair Housing Act. The City of Brandon, Mississippi is defendant in a case where the Homebuilders of Mississippi allege a new city ordinance has a disparate impact on minorities. The ordinance establishes a minimum for the size of new homes built in the city and the claim is that the minimum causes prices to be too high for new homes, having a disparate impact on minorities. Worked for defendant. Deposition, August 2008. Case ongoing.

#### *Toxic Tort:*

Three separate **Toxic Tort** property value diminution cases filed in Florida between 1998 and the present. Deposition for the largest and latest case in 2001. All three cases were environmental contamination cases, with class actions brought against manufacturer. Worked for defense in all three cases on class certification issues and damages calculations. Deposed in last case, First Case class was not certified, Second case settled. Third: Bernice Samples, et al, v. Conoco, Inc.; Agrico Chemical Company; and Escambia Treating in the Circuit Court of the First Judicial Circuit in and for Escambia County, FL, Division: “J”, June 2002, Deposition; Case settled.

#### *Other Antitrust:*

North Jackson Pharmacy, Inc. et al v. Express Scripts, Inc. et al. Independent Pharmacies filed an antitrust case against Pharmacy Benefit Managers (PBMs). Worked for plaintiffs. Deposed in July, 2005; class certified.

North Jackson Pharmacy, Inc. et al v. Caremark Pharmacies filed an antitrust case against Pharmacy Benefit Managers (PBMs). Worked for plaintiffs. Deposed in May, 2006; class certification pending.

REBUTTAL REPORT  
 REVIEW OF PCA AND PROJECTABILITY OF BIOMARKER INFERENCES FROM THE ILLINOIS RIVER WATERSHED

---

*Other cases:*

Castro v. Ford Motor, Inc. **Wrongful Death** Suit filed in California. Deposition and Testimony in 2001. Worked for defendant. Survey used in case regarding use of Ford Explorers by the general public. Critiqued survey and damages calculations as rebuttal expert. Jury found in favor of Ford.

Mullinax v. Buffalo Rock. **Wrongful Death** Suit in Alabama. Deposition and Testimony in 2004. Worked for plaintiff. Sampling of trucks from Pepsi bottling plant taken and analyzed to demonstrate that Pepsi \ Buffalo Rock drivers frequently speed, even after plaintiffs mother was killed by speeding fully loaded truck. Results were that 70 to 80 percent of trucks were observed speeding during a three month period, and 90 percent of “roll-up” trucks were speeding during this period. Jury found in favor of plaintiff with sizable award.

BMW Management, Inc. v. Sizzler, Inc.. Lost value and population estimates for population affected in a marketing case where a franchisor allowed a new franchise to be built in the “blocked area” around an already existing franchise. Worked for plaintiff. Case settled – deposition, January 2006.

Silver Pines Homeowners Association et al. v. Silver Pine Builders et al. **Construction Defects** Damages case regarding the calculation of damages based on a sample of housing units inspected and resulting damages extrapolated to the full population of units built in a new subdivision. Worked for defense. Case settled – deposition, May 2007.

**Fair Labor Standards Act (FLSA)** case filed against Dollar General involving claims for overtime not paid for store managers. Analysis of hours worked, duties performed, activity types. Case initially resolved in Summary Judgment against plaintiffs. On appeal reestablished and awaiting trial. Deposition, July 2007.

DA-HEEM RODGERS, ET AL. V. AVERITT EXPRESS, INC., **Fair Labor Standards Act (FLSA) class action**, case involving claims for overtime not paid for truck drivers. Analysis based on travel patterns and frequency of involvement in interstate commerce and damages calculation. Worked for plaintiffs. Case ongoing – deposition, June, 2008.